Preprint ANL/MCS-P1964-1011

Noname manuscript No.
(will be inserted by the editor)

Higher-Order Confidence Intervals for Stochastic Programming using Bootstrapping

Mihai Anitescu · Cosmin G. Petra

Received: date / Accepted: date

Abstract We present a novel approach based on bootstrap for constructing confidence intervals for the optimal value of a stochastic programming problem. We propose a statistical estimator of the optimal value and prove under some regularity conditions of the stochastic problem that the probability coverage of the confidence intervals based on our estimator converges almost one order faster than the standard estimates. This feature allows the construction of accurate and reliable confidence intervals for applications affording only a small number of samples. The good convergence properties are demonstrated with two numerical examples.

Keywords Stochastic programming · Nonlinear programming · Bootstrap · Exponential convergence

Mathematics Subject Classification (2000) 90C15 · 90C30 · 62F40

1 Introduction

We explore the assymptotics of statistical estimates for stochastic programming problems of the form

$$\min_{x \in X} f(x) := \mathbb{E}F(x, u). \tag{1}$$

Here we assume that the feasible set X is defined by three-times continuously differentiable constraints functions

$$X := \{x \in K | g_i(x) = 0, i = 1, 2, \ldots, p;\ g_i(x) \leq 0, i = p+1, \ldots, q\}. \tag{2}$$

M. Anitescu · C. G. Petra
Mathematics and Computer Science Division, Argonne National Laboratory,
9700 South Cass Avenue, Building 240, Argonne, IL 60439, USA.
E-mail: anitescu,petra@mcs.anl.gov

We use u to represent a random variable over the probability space (Ω, μ, P) with values in $\mathbb{B}_u$, a bounded subset of $\mathbb{R}^m$.

We seek to understand the properties of approximations to the problem (1) that are brought about by sample average approximation (SAA) problem

$$\min_{x \in X} f^N(x), \tag{3}$$

where $f^N(x) = f^N(x, \omega) = \frac{1}{N} \sum_{i=1}^{N} F(x, u_i(\omega))$. Here u_i, $i = 1, 2, \ldots, N$, are identical independent and identically distributed (i.i.d.) realizations of u.

Statistical estimates for the stochastic program (1) are obtained from the SAA (3) [12] and based on convergence in distribution of the type

$$\tau_N \left(v_N - v^* \right) \xrightarrow{\mathcal{D}} V. \tag{4}$$

Here v^* is the quantity to estimate (for example, optimal value of a stochastic program (1)), v_N is the estimator that depends on the sample size N, τ_N is a normalizing sequence (for example, $\tau_N = \frac{\sqrt{N}}{\sigma_N}$, where σ_N is an estimate of the standard deviation of the estimator v_N), and V is a fixed distribution function such as the standard normal distribution function.

Such estimates are essential, for example, for the construction of confidence intervals and they are known to work well in the limit of large sample sizes. On the other hand, in an increasing set of applications the collection of samples is expensive. One example is the situation where the system subject to the optimization under uncertainty (for example, an energy system with massive penetration of renewable sources) is affected by complex spatio-temporal uncertainty [4]. In this case, samples from the distribution are produced by numerical weather prediction codes that can be enormously expensive; hence, cannot realistically afford a large sample size, and instead, most times has to be content with fewer than 100 samples. This situation raises questions about the suitability of assuming the asymptotic regime described above in constructing such estimates, since such statements by themselves allow in principle arbitrarily slow convergence rates.

To this end, we investigate the rate of convergence with sample size in statements such as (4) for statistical estimates connected to stochastic programming. For example we seek statements of the type

$$F_N(x) = F(x) + O(N^{-a}),$$

where $F_N(x)$ is the cumulative distribution function of $\tau_N \left(v_N - v^* \right)$ and $F(x)$ is the cumulative distribution function of V. Here $a > 0$ is the order of convergence. Moreover, we propose estimators that converge faster than do the typical choices, in the sense that their parameter a is larger. Such results offer the promise—which we demonstrate in this work—of more accurate probability statements about the quality of the estimator even at low sample sizes.

We will obtain faster estimators by means of a classical technique: bootstrapping. However, bootstrap theory works well only for estimators that are smooth functions of means and other low-order moments of a random variables

[9]. This situation, as we indicate later, cannot be assumed for most stochastic programming problems. Therefore an important technical challenge, which we solve here, is to develop analytical tools for bootstrap estimates in stochastic programming.

2 Confidence Intervals

A confidence interval (CI) is a set of possible values of a statistical parameter θ of a random variable $\mathbf{X}$. The α-confidence interval for θ is any random interval $[L(\mathbf{X}), U(\mathbf{X})]$ that satisfies

$$P(\theta \in [L(\mathbf{X}), U(\mathbf{X})]) = \alpha, \tag{5}$$

where $\alpha \in (0, 1)$ is the confidence level, or the coverage probability of the interval [3]. Most commonly used are the 90%- and 95%-confidence intervals, obtained for $\alpha = 0.90$ and $\alpha = 0.95$, respectively. An *equal-tailed* two-sided confidence interval is a CI of the form (5) satisfying

$$P(\theta \leq L(\mathbf{X})) = (1 - \alpha)/2 = P(\theta > U(\mathbf{X})), \tag{6}$$

that is, an interval with the same level probability in each tail. *Symmetric* two-sided CIs are constructed around a statistic $\hat{\theta}$ corresponding to the parameter θ and have the form

$$P(|\theta - \hat{\theta}| \leq \tilde{U}(\mathbf{X})) = \alpha. \tag{7}$$

One-sided intervals are the intervals of the form (5) for which $L(\mathbf{X}) = -\infty$.

Notation: For the rest of the paper we use the following notation: z_α denotes the quantiles of the standard normal distribution, being the solution of the equation $\alpha = \Phi(z_\alpha)$, where $\Phi(x) = \int_{-\infty}^{x} \phi(x)dx$ is the standard normal cumulative distribution function and $\phi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ is the standard normal probability density function.

In this paper we assume that the distribution function F of the population $\mathbf{X}$ is unknown and only an i.i.d sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ drawn from F is available. In constructing CIs, sample-based statistics $\hat{W}$ and $\hat{V}$ are chosen such that

$$\frac{\hat{W} - \theta}{\hat{V}} \to \mathcal{N}(0, 1), \text{ as } N \to \infty \tag{8}$$

holds based on the central limit theorem. $(\hat{W} - \theta)/\hat{V}$ is called a *pivotal* statistic because asymptotically it does not depend on the population parameters. The confidence interval is taken to be $[\hat{W} - z_{(1+\alpha)/2}\hat{V}, \hat{W} - z_{(1-\alpha)/2}\hat{V}]$; however, for a fixed N, the interval is only approximate since it is only asymptotically exact, that is,

$$\lim_{n\to\infty} P(\theta \in [\hat{W} - z_{(1+\alpha)/2}\hat{V}, \hat{W} + z_{(1-\alpha)/2}\hat{V}]) = \alpha. \tag{9}$$

For example, in computing a CI for the first moment of $\mathbf{X}$, $\theta = \mathbb{E}[\mathbf{X}]$, $\hat{W}$ can be taken to be the sample mean $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ and $\hat{V}$ to be $\hat{\sigma}/N^{1/2}$, where $\hat{\sigma}$

is the standard deviation of the sample defined by $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$. An approximate α-level CI can be taken to be $[\bar{X} - N^{-1/2}z_{(1+\alpha)/2}\hat{\sigma}, \bar{X} - N^{-1/2}z_{(1-\alpha)/2}\hat{\sigma}]$.

For a given N, the coverage error is defined to be the nominal coverage α minus the approximate coverage $P(\theta \in [\hat{W} - z_{\alpha/2}\hat{V}, \hat{W} + z_{\alpha/2}\hat{V}])$. The concept of correctness of a CI characterizes the rate of convergence of the coverage error to zero with N. For example, it holds that

$$P(\theta \in [\hat{W} - z_{(1+\alpha)/2}\hat{V}, \hat{W} - z_{(1-\alpha)/2}\hat{V}]) = \alpha + O(N^{-1/2}), \tag{10}$$

see [9]. In general, a CI $[L, U]$ with L and U depending on N is said to be k-order correct ($k \geq 1$ being an integer) if

$$P(\theta \in [L, U]) = \alpha + O(N^{-k/2}). \tag{11}$$

Higher order CIs can be obtained based on the theory of Edgeworth expansions, a refinement of (8) that develop $(\hat{W} - \theta)/\hat{V}$ as a series of powers of $N^{-1/2}$. The connection with bootstrap is presented in Section 3.1.

Asymptotics of Confidence Intervals

To discuss some of the difficulties in the asymptotics of confidence intervals, we review some of the convergence concepts concerning a random variable over a probability space $(\Omega, \mathcal{F}, P)$. The core concepts in this subsection can be found in [1,12].

We recall that a random variable is a P-measurable mapping from Ω to $\mathbb{R}^d$ for some d. For sequences of random variables $\mathbf{X}_N(\omega), \omega \in \Omega$, we encounter the following convergence concepts relative to a random variable $\mathbf{X}(\omega)$. We recall that the probability of a statement is the measure of the set for which the statement is true with respect to the probability measure P. For example $P(\mathbf{X}_1 > \mathbf{X}_2) = P(\{\omega | \mathbf{X}_1(\omega) > \mathbf{X}_2(\omega)\})$.

- *Convergence in probability:* The sequence of random variables $\mathbf{X}_N(\omega)$ converges in probability to the random variable $\mathbf{X}(\omega)$ if

$$\forall \epsilon > 0, \quad \lim_{N\to\infty} P(|\mathbf{X}_N - \mathbf{X}| \geq \epsilon) = 0. \tag{12}$$

- *Convergence with probability 1 (almost sure convergence):* The sequence of random variables $\mathbf{X}_N(\omega)$ converges to the random variable $\mathbf{X}(\omega)$ with probability 1 (almost surely) if

$$P\left(\lim_{N\to\infty} |\mathbf{X}_N - \mathbf{X}| = 0\right) = 1. \tag{13}$$

- *Convergence in distribution:* The sequence of random variables $\mathbf{X}_N(\omega)$ converges to the random variable $\mathbf{X}(\omega)$ in distribution if

$$\lim_{N\to\infty} F_N(x) = F(x), \quad \text{at all continuity points of } F(x). \tag{14}$$

  where $F_N(x)$ are the cumulative distributions of $\mathbf{X}_N(\omega)$ and $F(x)$ is the cumulative distribution of $\mathbf{X}(\omega)$. Here, for a random variable $Y(\omega)$, we define the cumulative distribution F_Y to be

$$F(y) = P\left(Y^1 \leq y^1, Y^2 \leq y^2, \ldots, Y^d \leq y^d\right), \quad \forall y \in \mathbb{R}^d. \tag{15}$$

We also have that convergence with probability 1 is the strongest of the three. It implies convergence in probability, which in turn implies convergence in distribution [1, Theorem 25.2].

As suggested in (11), we are interested in making statements about the asymptotics of the coverage intervals. That is, we would like to make statements of the type (for example, for real-valued random variables)

$$P(\mathbf{X}_N \leq x) - P(\mathbf{X} \leq x) = O(N^{-b}), \tag{16}$$

for some positive number b and at all x and where $O(\cdot)$ is the Landau asymptotic notation. Here $\mathbf{X}(\omega)$ may be the target random variable and $\mathbf{X}_N(\omega)$ an approximation.

In stochastic programming, asymptotics *in distribution* or *in probability* are constructed by means of the delta theorem [12, Theorem 7.59]. Such asymptotics have the flavor, for example, of

$$\mathbf{X}_N(\omega) - \mathbf{X}(\omega) = o_P(N^{-a}), \tag{17}$$

taken to mean that $N^a \left|\mathbf{X}_N - \mathbf{X}\right|$ converge to 0 in probability.

Unfortunately, such asymptotics in probability do not translate to similar asymptotics in coverage, of the type (16). This difficulty is also alluded to in the development of the asymptotics for classical bootstrapping [9].

Indeed, consider the following random variables over the probability space given by the unit interval, with the usual Lebesgue measure:

$$\mathbf{X}(\omega) = \omega, \quad \mathbf{X}_N(\omega) = \begin{cases} -1, & 0 \leq \omega < \frac{1}{\log(N+1)}, \\ \omega, & \frac{1}{\log(N+1)} \leq \omega \leq 1. \end{cases} \tag{18}$$

It immediately follows that $P(N^a \left|\mathbf{X}_N - \mathbf{X}\right| = 0) \geq 1 - \frac{1}{\log(N+1)}$ and thus

$$P(\lim_{N\to\infty} N^a \left|\mathbf{X}_N - \mathbf{X}\right| = 0) = 1,$$

for any positive a. This means that $N^a \left|\mathbf{X}_N - \mathbf{X}\right|$ converges almost surely, and, thus, in probability, to 0 for any a positive, irrespective of how large.

On the other hand, we have that $P(\mathbf{X} \leq 0) = 0$, but $P(\mathbf{X}_N \leq 0) = \frac{1}{\log(N+1)}$. Therefore, clearly, in this case, (16) cannot be satisfied for any positive b, *irrespective of how small*.

We thus conclude that the convergence in probability asymptotics of the type provided by the delta theorem are insufficient to obtain similar asymptotics in coverage. We thus take a different technical direction, based on large deviation theories for measures with compact support.

3 Bootstrap Confidence Intervals

Most concepts and discussions present in this section are taken from [8,7,9]. Bootstrap sampling refers to the use of same-size samples $\mathcal{X}^* = \{X_1^*, \ldots, X_N^*\}$ drawn repeatedly with replacement from the original sample $\mathcal{X} = \{X_1, \ldots, X_N\}$. This is equivalent to saying that $\mathcal{X}^*$ is drawn from the empirical distribution function $\hat{F}$ of $\mathcal{X}$. Observe that the number of distinct bootstrap samples is finite although very large even for relatively small N and is given by the binomial coefficient $\binom{2N-1}{N} = \frac{(2N-1)!}{N!\,(N-1)!}$. A bootstrap replication θ^{*i} is obtained by evaluating $\hat{\theta}$ based on a bootstrap sample $\mathcal{X}_i^*$ instead of the original sample $\mathcal{X}$, $i = 1, \ldots, \binom{2N-1}{N}$. We denote by θ^* the random variable that is obtained from evaluating $\hat{\theta}$ based using bootstrap samples $\mathcal{X}_i^*$. Observe that θ^{*i} are i.i.d realizations of θ^*.

Bootstrap replications mimic replications $\hat{\theta}$ that would have been obtained by sampling the true distribution F. The simple and yet powerful idea behind bootstrap is that it samples an approximating distribution, for example, the empirical distribution $\hat{F}$, when the true population distribution is not available. The advantage of bootstrapping is that the bootstrap distribution which we denote by F^* is known and can be easily sampled; and, since it is finite, any parameters depending on it can be worked out to arbitrary accuracy by using simulation.

Basic Bootstrap Confidence Intervals

Bootstrap CIs are obtained by solving bootstrap approximations of equation (5). One such variant, known as *bootstrap percentile* CI, is computed by simply considering the bootstrap distribution F^* instead of the F in (6), and solving for the $(1-\alpha)/2 \cdot 100$th and $(1+\alpha)/2 \cdot 100$th percentiles θ_{L^*} and θ_{U^*} of the bootstrap distribution from

$$P(\theta^* \leq \theta_{L^*}) = (1-\alpha)/2 = P(\theta^* > \theta_{U^*}). \tag{19}$$

Bootstrap percentile CIs are affected by the bias of $\hat{\theta}$, especially for small sample sizes. BCa (bias-corrected and accelerated) intervals presented in Section 3.1 are of this type, but they explicitly correct the bias (and the skewness) and are second-order correct.

An equivalent formulation of the equation (6) is

$$P(\theta \leq \hat{\theta} - \theta_L) = (1-\alpha)/2 = P(\theta > \hat{\theta} + \theta_U), \tag{20}$$

which imply that $[\hat{\theta} - \theta_L, \hat{\theta} + \theta_U]$ is an α-level equal-tailed two-sided CI for θ. The bootstrap variant of the above equation proposed in [9] is

$$P(\hat{\theta} \leq \theta^* - \hat{\theta}_L) = (1-\alpha)/2 = P(\hat{\theta} > \theta^* + \hat{\theta}_U), \tag{21}$$

and is obtained in the same idea of replacing F with F^*. The bootstrap CI is taken to be $[\theta^* - \hat{\theta}_L, \theta^* + \hat{\theta}_U]$. Bootstrap-t CIs presented in Section 3.1 are computed on this principle but they use pivotal statistics of the form (8) and achieve an extra order of correctness (*i.e.*, second-order correctness) over $[\theta^* - \hat{\theta}_L, \theta^* + \hat{\theta}_U]$ bootstrap CI.

3.1 High Order Bootstrap Confidence Intervals

Consider the random vector $\mathbf{X}$ and the parameter $\theta = f(\mu)$, where f is a smooth function and $\mu = \mathbb{E}[\mathbf{X}]$. We note that any parameter defined as a smooth function of the moments of $\mathbf{X}$ fits this setup; see [9, Section 2.4].

3.1.1 Hall CIs

Given an i.i.d. sample $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$ withdrawn from $\mathbf{X}$, the parameter θ is estimated by the statistic $\hat{\theta} = f(\bar{X})$. Also let $\sigma^2 = h^2(\mu)$ be the asymptotical variance of $N^{1/2}\hat{\theta}$, with h being a known smooth function. Usually σ^2 is not known because μ is not available, and an estimate $\hat{\sigma}^2 = h^2(\bar{X})$ is used instead. Denote by $H(x)$ and $K(x)$ the distribution functions of $N^{1/2}(\hat{\theta}-\theta)/\sigma$ and $N^{1/2}(\hat{\theta} - \theta)/\hat{\sigma}$, respectively. The α-level quantiles $x_\alpha = H^{-1}(\alpha)$ and $y_\alpha = K^{-1}(\alpha)$ can be used to construct exact confidence intervals for $\hat{\theta}$. More specifically, one can easily verify that

$$P(\theta \le \hat{\theta} - N^{-1/2}\sigma x_{1-\alpha}) = P(\theta \le \hat{\theta} - N^{-1/2}\hat{\sigma} y_{1-\alpha}) = \alpha \tag{22}$$

and that one-sided intervals

$$I_1 = I_1(\alpha) = \left(-\infty, \hat{\theta} - N^{-1/2}\sigma x_{1-\alpha}\right) \tag{23}$$

$$J_1 = J_1(\alpha) = \left(-\infty, \hat{\theta} - N^{-1/2}\hat{\sigma} y_{1-\alpha}\right) \tag{24}$$

are exact α-level confidence intervals for θ.

Analogous α-level equal-tailed two-sided CIs can be obtained based on one-sided intervals defined above:

$$I_2 = I_1\left(\frac{1+\alpha}{2}\right) \setminus I_1\left(\frac{1-\alpha}{2}\right) = \left(\hat{\theta} - N^{-1/2}\sigma x_{\frac{1+\alpha}{2}}, \hat{\theta} - N^{-1/2}\sigma x_{\frac{1-\alpha}{2}}\right), \tag{25}$$

$$J_2 = J_1\left(\frac{1+\alpha}{2}\right) \setminus J_1\left(\frac{1-\alpha}{2}\right) = \left(\hat{\theta} - N^{-1/2}\hat{\sigma} y_{\frac{1+\alpha}{2}}, \hat{\theta} - N^{-1/2}\hat{\sigma} y_{\frac{1-\alpha}{2}}\right). \tag{26}$$

We adopt the terminology from [9] and call J intervals *percentile-t* confidence intervals. The same term is used, for example, in [5], for CIs built on the Student t-statistic; however, in the present work it refers only to J intervals.

Since the quantiles x_α and y_α are not available (θ is not a priori known) one has to rely on approximations of x_α and y_α to build confidence intervals. One such approximation is given by the central limit theorem which states that under some regularity conditions, $N^{1/2}(\hat{\theta} - \theta)/\sigma$ and $N^{1/2}(\hat{\theta} - \theta)/\hat{\sigma}$ are asymptotically normally distributed and that x_α and y_α tend to $z_\alpha = \Phi^{-1}(\alpha)$ as $N \to \infty$ [3]. The corresponding"normal approximation" CIs are computed by replacing x_α (or y_α) with z_α for any N. Such intervals are only first-order correct, for reasons presented below, and one has to rely on bootstrapping to obtain higher-order correct CIs. The benefit of bootstrapping is that it considers additional terms from the Edgeworth expansions of the distribution

functions of $N^{1/2}(\hat{\theta}-\theta)/\sigma$ and $N^{1/2}(\hat{\theta}-\theta)/\hat{\sigma}$ in the approximation of x_α and y_α.

The bootstrap CIs approximating the true CIs (25) and (26) are obtained by bootstrapping the quantiles x_α and y_α. Consider the bootstrap counterparts $N^{1/2}(\theta^*-\hat{\theta})/\hat{\sigma}$ and $N^{1/2}(\theta^*-\hat{\theta})/\sigma^*$ of $N^{1/2}(\hat{\theta}-\theta)/\sigma$ and $N^{1/2}(\hat{\theta}-\theta)/\hat{\sigma}$, and denote by $\hat{H}(x)$ and $\hat{K}(x)$ their distribution functions. Here the bootstrap estimate of the variance is defined based on the same principle as θ^* was defined in the previous paragraph, that is, $\sigma^{*2} = h^2(\bar{X}^*)$. The bootstrap approximations to α-level quantiles x_α and y_α are taken to be $\hat{x}_\alpha = \hat{H}^{-1}(\alpha)$ and $\hat{y}_\alpha = \hat{K}^{-1}(\alpha)$. Since the distributions $\hat{H}(x)$ and $\hat{K}(x)$ are discrete, $\hat{x}_\alpha$ and $\hat{y}_\alpha$ are precisely defined by

$$\hat{x}_\alpha = \inf\{x|P[N^{1/2}(\theta^*-\hat{\theta})/\hat{\sigma} \le x|\hat{F}] \ge \alpha\}, \tag{27}$$

$$\hat{y}_\alpha = \inf\{y|P[N^{1/2}(\theta^*-\hat{\theta})/\sigma^* \le y|\hat{F}] \ge \alpha\}. \tag{28}$$

The sample-based quantiles $\hat{x}_\alpha$ and $\hat{y}_\alpha$ can be computed to arbitrary precision by sampling $\hat{F}$. The one-sided bootstrap percentile-t CIs corresponding to the true CIs (23) and (24) are defined to be

$$\hat{I}_1 = \left(-\infty, \hat{\theta} - N^{-1/2}\hat{\sigma}\hat{x}_{1-\alpha}\right), \tag{29}$$

$$\hat{J}_1 = \left(-\infty, \hat{\theta} - N^{-1/2}\hat{\sigma}\hat{y}_{1-\alpha}\right). \tag{30}$$

The definition of two-sided bootstrap CIs I_2 and J_2 is identical to (25) and (26) with $\hat{I}_1$ and $\hat{J}_1$ replacing I_1 and J_1, respectively.

The dissimilarity between I_1 and J_1 comes from the use of different variance estimates, namely, an unknown quantity $\sigma^2 = h^2(\mu)$ for I_1 and a computable quantity $\hat{\sigma}^2 = h^2(\bar{X})$ for J_1. There are two approximations when bootstrapping I_1, x_α by $\hat{x}_\alpha$ and σ by $\hat{\sigma}$; however, there is only one approximation when bootstrapping J_1, y_α by $\hat{y}_\alpha$. This aspect turns out to be crucial in the order of correctness of the bootstrap CIs intervals $\hat{I}_1$ and $\hat{J}_1$: $\hat{I}_1$ is only first-order correct and $\hat{J}_1$ is second-order correct.

The correctness analysis of $\hat{I}_1$ and $\hat{J}_1$ is based on Cornish-Fisher expansions of $\hat{x}_\alpha$ and $\hat{y}_\alpha$. These expansions are inverted Edgeworth expansions of the distribution function of $H(x)$ and $K(x)$ (and their bootstrap counterparts). Edgeworth expansions improve the central limit theorem in at least two aspects. First, they provide additional terms in the approximation; second, they are true asymptotical expansions in the sense that the error is controlled. More specifically, under regularity conditions on the moments of $\mathbf{X}$ and continuity of the first $(k+2)$th derivatives of f, it can be proved ([8, Section 1.4 and Section 5], also [9, Theorem 2.2]) that

$$H(x) = \Phi(x) + N^{-1/2}p_1(x)\phi(x) + \ldots + N^{-k/2}p_k(x)\phi(x) + O(N^{-(k+1)/2}) \tag{31}$$

holds uniformly in x, where $p_i(x)$ is a polynomial of degree $i+1$ whose coefficients depend on the first $i+2$ derivatives of f and first $i+2$ moments of $\mathbf{X}$. Under the same assumptions, a similar Edgeworth expansion holds for $K(x)$:

$$K(x) = \Phi(x) + N^{-1/2}q_1(x)\phi(x) + \ldots + N^{-k/2}q_k(x)\phi(x) + O(N^{-(k+1)/2}), \tag{32}$$

where $q_i(x)$ are polynomials that have the same properties and are computed exactly as the polynomials $p_i(x)$ from (31) with $\hat{\sigma}$ replacing σ.

Edgeworth expansions can be "inverted" to be obtain Cornish-Fisher expansions of the quantiles of a distribution function. Cornish-Fisher expansions are also asymptotic series that hold uniformly in $\varepsilon < \alpha < 1 - \varepsilon$, for any $\varepsilon \in (0, 1/2)$. In the context of CIs, Cornish-Fisher expansions are known as a tool that corrects the effects of non-normality. Under the same assumptions under which Edgeworth expansion (31) of $H(x)$ or (32) of $K(x)$ exists, one can be prove (see [9, Theorem 2.4], see also [7]) that

$$x_\alpha = z_\alpha + N^{-1/2} p_{11}(z_\alpha) + \ldots + N^{-k/2} p_{k1}(z_\alpha) + O(N^{-(k+1)/2}), \quad (33)$$
$$y_\alpha = z_\alpha + N^{-1/2} q_{11}(z_\alpha) + \ldots + N^{-k/2} q_{k1}(z_\alpha) + O(N^{-(k+1)/2}), \quad (34)$$

uniformly in $\varepsilon < \alpha < 1 - \varepsilon$, for any $\varepsilon \in (0, 1/2)$. Here the polynomials p_{j1} and q_{j1} are of degree at most $j + 1$, odd for even j and even for odd j, and depend on cumulants of order up to $j + 2$.

The bootstrapped distribution function $\hat{H}(x)$ and $\hat{K}(x)$ also possess Edgeworth expansions which can be "inverted" to obtain Cornish-Fisher expansions of $\hat{x}_\alpha$ and $\hat{y}_\alpha$. These expansions have the form

$$\hat{x}_\alpha = z_\alpha + N^{-1/2} \hat{p}_{11}(z_\alpha) + \ldots + N^{-k/2} \hat{p}_{k1}(z_\alpha) + O(N^{-(k+1)/2}), \quad (35)$$
$$\hat{y}_\alpha = z_\alpha + N^{-1/2} \hat{q}_{11}(z_\alpha) + \ldots + N^{-k/2} \hat{q}_{k1}(z_\alpha) + O(N^{-(k+1)/2}). \quad (36)$$

The polynomials $\hat{p}_{k1}$ and $\hat{q}_{k1}$ are computed in the same way the polynomials p_{k1} and q_{k1} are computed, the only difference being that the sample moments replace the population moments. The last remark causes

$$\hat{p}_{k1} = p_{k1} + O_p(N^{-1/2}) \text{ and } \hat{q}_{k1} = q_{k1} + O_p(N^{-1/2}).$$

Observe that the difference between the polynomials is characterized by "order in probability". A random variable Z_N is said to have order δ_N in probability, written $Z_N = O(\delta_N)$, if the sequence

$$\lim_{t \to \infty} \limsup_{N \to \infty} P(|Z_N / \delta_N| > t) = 0.$$

Consequently, the bootstrap quantiles are second-order correct to the true quantiles since

$$\hat{x}_\alpha - x_\alpha = N^{-1/2}(\hat{p}_{11}(z_\alpha) - p_{11}(z_\alpha)) + O_p(N^{-1}) = O_p(N^{-1}), \quad (37)$$
$$\hat{y}_\alpha - y_\alpha = N^{-1/2}(\hat{q}_{11}(z_\alpha) - q_{11}(z_\alpha)) + O_p(N^{-1}) = O_p(N^{-1}). \quad (38)$$

The second-order correctness of the bootstrap quantiles gives benefits only in the case of bootstrap percentile-t intervals J. To see this, we first compare the endpoints of $\hat{I}_1$ with I_1:

$$\begin{aligned}(\hat{\theta} - N^{-1/2}\hat{\sigma}\hat{x}_{1-\alpha}) - (\hat{\theta} - N^{-1/2}\sigma x_{1-\alpha}) &= N^{-1/2} x_{1-\alpha}(\hat{\sigma} - \sigma) + O_p(N^{-3/2}) \\ &= N^{-1} x_{1-\alpha} N^{1/2}(\hat{\sigma} - \sigma) + O_p(N^{-3/2}) = O_p(N^{-1}),\end{aligned}$$

since $N^{1/2}(\hat{\sigma}-\sigma) = O_p(1)$. On the other hand, the endpoints of $\hat{J}_1$ and J_1 satisfy

$$\begin{aligned}(\hat{\theta} - N^{-1/2}\hat{\sigma}\hat{y}_{1-\alpha}) - (\hat{\theta} - N^{-1/2}\hat{\sigma}y_{1-\alpha}) &= N^{-1/2}\hat{\sigma}y_{1-\alpha}(\hat{\sigma}-\hat{\sigma}) + O_p(N^{-3/2}) \\ &= O_p(N^{-3/2}).\end{aligned}$$

Therefore, the endpoints of $\hat{J}_1$ have an extra order of correctness over the endpoints of $\hat{I}_1$. This shows the advantage of using the pivotal statistics $N^{1/2}(\hat{\theta}-\theta)/\hat{\sigma}$ over the nonpivotal statistic $N^{1/2}(\hat{\theta}-\theta)/\sigma$ in computing confidence intervals.

$\hat{J}_1$ is not obviously second-order correct as per definition (11), because, as discussed in §2, asymptotic convergence in probability does not imply asymptotics in coverage. In fact, the analysis of the size of coverage error for one-sided bootstrap confidence intervals is elaborate. An Edgeworth expansion of the coverage probabilities (see [9, Proposition 3.1 and Theorem 5.3]) is used to show

$$P(\theta \in \hat{I}_1(\alpha)) = \alpha + O(N^{-1/2}), \tag{39}$$

$$P(\theta \in \hat{J}_1(\alpha)) = \alpha + O(N^{-1}), \tag{40}$$

uniformly with $\varepsilon < \alpha < 1-\varepsilon$, for any $\varepsilon \in (0, \frac{1}{2})$.

3.1.2 Bias-corrected and accelerated bootstrap CIs

The bias-corrected and accelerated (BCa) bootstrap methods introduced by Efron [5] do not use a pivotal statistics and work directly in the bootstrap distribution F^*.

Define $G(x) = P(\hat{\theta} \le x)$ and $\hat{G}(x) = P(\theta^* \le x)$, the cumulative distribution functions of $\hat{\theta}$ and θ^*. The computation of the bootstrap CI based on (19) reduces to computing the $(1-\alpha)/2$- and $(1+\alpha)/2$-level quantiles of $\hat{G}$, namely $\hat{v}_{(1-\alpha)/2} = \hat{G}^{-1}((1-\alpha)/2)$ for θ_{L^*} and $\hat{v}_{(1+\alpha)/2} = \hat{G}^{-1}((1+\alpha)/2)$ for θ_{U^*}. Bias occurs mainly because of lack of symmetry of the probability distribution of $\hat{\theta}$. It is corrected by shifting the quantiles $\hat{v}_\alpha = \hat{G}^{-1}(\alpha) = \hat{G}^{-1}(\Phi(z_\alpha))$ to $\hat{v}_{BC,\alpha} = \hat{G}^{-1}(\Phi(z_\alpha + 2\hat{m}))$, where $\hat{m} = \Phi^{-1}(\hat{G}(\hat{\theta}))$ accounts for centering error occurring at the median.

A second adjustment for skewness yields the BCa quantile

$$\hat{v}_{BCa,\alpha} = \hat{G}\left(\Phi\left(\hat{m} + \frac{z_\alpha + \hat{m}}{1 - \hat{a}(z_\alpha + \hat{m})}\right)\right), \tag{41}$$

where $\hat{a}$ is known as the acceleration constant and approximates the skewness of (a first-order approximation to) the pivotal statistic $(\hat{\theta}-\theta)/\hat{\sigma}$ [6,9]. One of the most common expressions of $\hat{a}$ is

$$\hat{a} = N^{-1/2}\frac{1}{6}\frac{\hat{A}}{\hat{\sigma}^3}, \tag{42}$$

where

$$\hat{A} = \frac{3}{N}[f'(\bar{X})]^3 \sum_{i=1}^{N}(X_i - \bar{X})^3. \tag{43}$$

Analysis of the correctness order of the BCa intervals from [9, Chapter 3.10] reveals that the BCa quantile is second-order correct to $\hat{y}_\alpha$:

$$\hat{v}_{BCa,\alpha} = \hat{y}_\alpha + O(N^{-1}), \tag{44}$$

which implies that one-sided BCa intervals of the form

$$\hat{J}_{1,BCa}(\alpha) = (-\infty, \hat{v}_{BCa,\alpha}) \tag{45}$$

are second-order correct.

In practice, $\hat{a}$ is estimated by using the jackknife method. Jackknifing is a resampling technique that leaves out one or more observations from the sample when replicating the statistic, see [6, Chapter 14] for a thorough discussion. The computational details of the jackknife are given in Section 5.2.

4 Asymptotic Results for Stochastic Programs

This section is close to [12, Theorem 5.8] in its aims, with a couple of significant differences. First, we move away from convergence in probability and we seek exponential convergence results, given the limited usability of convergence in probability results for asymptotics of coverage intervals, as pointed out in §2. Second, the nature of the analytical results in [12, Theorem 5.8] provides the expansion of the optimal value of the SAA approximation to the stochastic program in terms of quantities evaluated at x^*, the solution of the stochastic program. We proceed the other way around: we approximate the optimal value of the stochastic program in terms of quantities evaluated at the solution of the sample average approximation. In turn, this will allow for the construction of higher-order estimates of the value of the stochastic program itself.

We consider the stochastic programs (1) whose objective function is an expectation, that is,

$$f(x) = E_u F(x, u) \tag{46}$$

and F is a sufficiently regular function whose properties will be described later. In addition, we assume that f itself can be computed only by evaluating the integral defining the expectation in (46).

Notation: We will use n for the dimension of the variable space x, m for the dimension of the range of the random variable u, and N for the number of samples.

We seek to understand the properties of approximations to the function f that are brought about by sample average approximation

$$f^N(x) = f^N(x, \omega) = \frac{1}{N}\sum_{i=1}^{N} F(x, u_i(\omega)) \tag{47}$$

and how these influence the relationship between minimizing f over a set X and minimizing f^N.

What makes reasoning about resampling applied to (1) difficult is that some of the most convenient tools for bootstrap analysis require that the target estimator be a smooth function of the expected value (or low-order moments) of $u(\omega)$ [9]. This is not the case in (1), where allowing for arbitrary nonlinear dependence with u in (47) does not result in the optimal value (when using f as objective function in minimizing over a set X, (1)) being a function of only a finite set of moments of u. It is conceivable that an approximation approach based on Taylor series followed by invoking results from [9] would work, but we want to explore the consequences of limited differentiability of F in u, so the latter approach would be inconvenient here.

We thus follow a different approach. We divide into two parts the problem of constructing estimators for the solution of optimization problems with objective function f. First, using large deviation theory, we construct estimates whose flavor is

$$\mathbb{P}(|N^b(\hat{\theta} - \theta)| > \epsilon) = f_1(\epsilon)N^{f_2(\epsilon)} \exp(-f_3(\epsilon)N^c).$$

Here θ is the quantity we aim to estimate (such as the optimal value of a program involving f) and $\hat{\theta}$ is the estimator. The quantities b, c are positive exponents, and f_1,f_2, and f_3 are positive-valued functions. Note the difference with concepts of convergence in probability, where the right-hand side would only be required to converge to 0, whereas here it converges *exponentially* to 0 with increasing N and fixed ϵ. Subsequently, since the estimator $\hat{\theta}$ contains only means of f and its derivatives with respect to x at a fixed point $\hat{x}$, bootstrap theory can be applied.

In this entire work, we make the following assumption:

[A0] $u(\omega)$ is a random variable whose image lies in a compact set $\mathbb{B}_u \subset \mathbb{R}^m$.

In addition, in this section we make two further assumptions that are important for establishing such convergence properties. That is, we assume that there exists a compact set K such that the following are satisfied.

[A1] $\nabla_x F(x, u)$ is directionally differentiable for any $x \in K$ and u in $\mathbb{B}_u$ with the directional derivatives uniformly bounded over $K \times \mathbb{B}_u$.

[A2] $f(x) = E_u F(x, u)$ is twice continuously differentiable in x over K, with Lipschitz continuous second derivative.

Discussion Assumption [A2] will be the most limiting assumption of our analysis. In particular, it does not allow us, in the current form of the theory, to apply our results to two-stage stochastic programming with inequality constraints in the second stage. Using results in [2], one can show, that, in the case of nonlinear stochastic programming where the second-stage problem is uniformly regular in terms of both first- and second-order conditions, [A1] will hold. Intuitively, it would seem that for some problem classes, [A2] would hold as well.

Indeed, consider the case $F(x,u) = \mathrm{sgn}(x-u)\,(x-u)^2$ and u is uniformly distributed in $[-1,1]$. It then follows that

$$\begin{aligned}\nabla_x f(x) &= \int_{-1}^{1} |x-u| du = \int_{-1}^{x} (-u+x)\, du + \int_{x}^{1} (x-u)\, dt \\ &= \frac{1}{2}(1+x)^2 + \frac{1}{2}(1-x)^2,\end{aligned}$$

whenever $x \in [-1,1]$. Hence, the sample average approximation may be not twice differentiable even though the average itself is (actually, infinitely differentiable).

Unfortunately, we have not succeeded in following this line of thought to have [A2] hold for a sufficiently significant class of two-stage stochastic programs with inequality constraints. For the case of three times differentiable equality constraints in the second stage, assumptions [A1] and [A2] seem to hold for the broad problem class, and thus they will hold for the case of smoothing the inequality constraints as brought about by, for example, using interior point with a fixed barrier parameter in the second stage.

In any case, two-stage stochastic programming with inequality constraints would need more analysis. We will thus content ourselves with assuming [A1] and [A2] in the more abstract framework defined by (46) (and, hence, (1)) and gauge their consequences.

Under [A0]–[A2] we have the following result.

Lemma 1 *There exist C_1^g and C_2^g such that*

$$\mathbb{P}\left(\sup_{x\in K} \left|\left|\nabla_x f^N(x) - \nabla_x f(x)\right|\right|_\infty \geq \epsilon\right) \leq C_1^g \frac{1}{\epsilon^n} \exp(-C_2^g N\epsilon^2).$$

Proof The result follows from [12, Theorem 7.67] applied componentwise to gradients. That result applies because the conditions leading to it are satisfied: uniform Lipschitz continuity of the gradient of F follows from [A1], and all moment generating conditions follow from [A0]. □

4.1 Exponential Convergence to a Neighborhood of the Solution

After considering the results for the approximation of the objective function $f(x) = \mathbb{E}F(x,u)$ by sample average approximation, we now concentrate on the exponential convergence property for the stochastic program (1). In addition, we assume that $X \subset K$ holds, where K is the compact set used in the preceding section to analyze the large deviations.

We make the following new assumptions about the solution of the stochastic program (1) and its SAA approximation (3):

[A3] The problem (1) has a unique solution x^* that is in the interior of K.

We then have that the probability that a solution x^N of (3) is outside any given neighborhood of x^* converges exponentially to 0 with increasing sample size. Formally, the statement is as follows.

Theorem 1 *Assume that [A1] and [A3] holds. Let $\epsilon > 0$ be such that $\mathcal{B}(x^*, \epsilon) \subset K$. Then, for any solution x^N of (3) we have that there exist $C_1^K, C_2^K > 0$ such that*

$$\mathbb{P}(\left|\left|x^N - x^*\right|\right| > \epsilon) < C_1^K \exp(-NC_2^K). \tag{48}$$

Proof The result follows from Corollary 5.19 in [12].

We verify that the conditions leading to that result do hold. Assumption [A1] ensures the uniform Lipschitz property of $F(x, u)$ and thus assumption (M5) in that result's assumptions hold. Moreover, as a result of Hoeffding's inequality (see the discussion following Corollary 5.19 in [12]), the moment-generating function of the random variable $Y = F(x, u) - f(x) - [F(x', u) - f(x)]$ is bounded above by $\exp(2L^2 \left|\left|x' - x\right|\right|^2 t^2)$ where L is a bound on the Lipschitz parameter. In turn, this implies that assumption (M6) used in Corollary 5.19 in [12] does hold. Assumption (M1) invoked in that reference holds immediately due to our assumption that u has a bounded range and that $X \subset K$, a compact set.

We can thus use those results to get the conclusion, where C_1^K and C_2^K are parameters depending on n, ϵ, L, and the diameter K. The proof is complete. □

4.2 A Locally Equivalent Unconstrained Optimization Program

The main aim of this section is to obtain a high-quality estimate of the value of the stochastic program (1) as a function of the sample size, N. To that end, we need to introduce a few assumptions about the quality of the solution point x^* of (1).

First, we introduce the Lagrangian of (1):

$$L(x, \lambda) = f(x) + \sum_{i=1}^{q} \lambda_i g_i(x).$$

At a feasible point x, we define by $I(x)$ the active set, the set of indexes of inequality constraints that are 0 (active) at x^*, that is,

$$I(x) := \{i | g_i(x) = 0, \quad i = p+1, p+2, \dots q\}.$$

The assumptions required are as follows.

[A4] At the point x^*, (1) satisfies the linear independent constraint qualification.
[A5] The unique Lagrange multiplier obtained as a result of [A4] satisfies the strict complementarity condition.
[A6] At the point x^*, (1) satisfies the second-order sufficient condition.

For completeness, we recall that the linear independence constraint qualification (LICQ) requires that the matrix of the active constraints

$$J(x) = \nabla g_{\mathcal{I}(x)}(x), \text{ where } \mathcal{I}(x) = \{1, 2, \dots, p\} \cup I(x), \tag{49}$$

be full row rank at x^*. Another way to state it is that

$$d \in \mathbb{R}^{\text{card}(\mathcal{I}(x^*))} \Rightarrow \left|\left|J(x^*)^T d\right|\right| \geq \sigma_J \left|\left|d\right|\right|. \tag{50}$$

Here, σ_J is the smallest singular value of $J(x^*)$, and by LICQ it must be a positive quantity. We define

$$J^*(x) = \nabla g_{\mathcal{I}(x^*)}(x). \tag{51}$$

Also, a Lagrange multiplier of (1) at x^* is a vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_q)$ satisfying

$$\nabla_x L(x^*, \lambda) = 0,\ \lambda_i \geq 0,\ i = p+1, \ldots, q,\ g_i(x^*) > 0 \Rightarrow \lambda_i = 0,\ i = p+1, \ldots, q.$$

As a consequence of the constraint qualification [A4], such a multiplier exists and is unique [11].

Strict complementarity implies that all Lagrange multipliers of the active inequality constraints are positive, that is,

$$\lambda^i > 0, \quad \forall i \in \mathcal{I}(x^*).$$

The second-order sufficient condition (under the strict complementarity condition) implies that there exists a $\sigma_L > 0$ such that

$$\nabla_i g_i(x^*) d = 0 \forall i \in \mathcal{I}(x^*) \Rightarrow d^T \nabla^2_{xx} L(x^*, \lambda) d \geq \sigma_L d^T d. \tag{52}$$

We also define a neighborhood of x^* wherein any solution of (1) with some perturbation of the objective function has the same feasible constraints and active set. Formally the result is the following.

Lemma 2 *There exist $\rho_S > 0$ and δ_S such that $\forall x \in X$, $||x - x^*|| \leq \rho_S$. We then have the following:*

a) $g_i(x) < 0, \forall i \notin \mathcal{I}(x^*)$.

b) $d \in \mathbb{R}^{card(\mathcal{I}(x^*))} \Rightarrow \left|\left|J(x^*)^T d\right|\right| \geq \frac{\sigma_J}{2} \left|\left|d\right|\right|$.

c) $\forall \eta \in \mathbb{R}^n$ *such that* $||\eta|| \leq \delta_S$ *and* $\forall x \in X$ *such that* $||x - x^*|| \leq \rho_S$, *we have that*

$$P_{J^{*\perp}} \left(\nabla f(x) + \eta\right) = - \sum_{i \in \mathcal{I}(x^*)} \lambda_i(x, \eta) \nabla g_i(x),$$

Here $\lambda_i(x, \eta) > 0$, $i \in \mathcal{I}(x^*)$, are the unique scalars with this property for given $x \in X$, and η. Here $P_{J^{*\perp}}$ denotes the orthogonal projection on the space spanned by the columns of $J^*(x)^T$.

d) $\forall \eta \in \mathbb{R}^n$, $||\eta|| \leq \delta_S$, $\forall x \in X$ *such that* $||x - x^*|| \leq \rho_S$, *we have that*

$$\nabla_i g_i(x) d = 0, \forall i \in \mathcal{I}(x^*) \text{ implies } d^T \nabla^2_{xx} L(x, \lambda(x, \eta)) d \geq \frac{\sigma_L}{2} d^T d.$$

Note: *All statements are made with respect to the active set at x^*.*

Proof Item (a) follows immediately from the fact that the twice differentiability of the data of (1) implies that the inactive set of a nonlinear program is robust to perturbations. Item (b) follows from the continuity of the singular value of a matrix with respect to the matrix data. Item (c) follows from the KKT conditions at x^*, which in turn implies that

$$P_{J^{*\perp}}\left(\nabla f(x^*)\right) = \sum_{i \in \mathcal{I}(x^*)} \lambda_i \nabla g_i(x^*),$$

together with the strict complementarity assumption [A5] and the fact that the projection operator is continuous. For item (d) we note that the previous items imply that $\lambda_i(x,\eta)$ are continuous functions of x,η. In turn continuity of the eigenvalues of a matrix with respect to its entries applied to the projection of $\nabla^2_{xx} L(x,\lambda(x,u))$ onto the nullspace of kernel of $J^*(x)$, and the second-order sufficient condition (52) give the conclusion. □

An immediate consequence of Lemma 2 is that any sufficiently small perturbation of (1) (where the size of the perturbation is governed by δ_S in terms of the size of the gradient) can have as stationary points only points that have the same active set as x^* and that satisfy the strict complementarity condition.

In particular, we have the following result.

Lemma 3 *Let $\mathcal{I}(x^N)$ be the active set at x^N, the solution of* (3)*. There exist $C_1^I > 0$ and $C_2^I > 0$ such that*

$$\mathbb{P}\left(\mathcal{I}(x^N) \neq \mathcal{I}(x^*)\right) \leq C_1^I \exp(-C_2^I N).$$

Proof From Theorem 1 we have that there exist C_1^K and C_2^K such that

$$\mathbb{P}\left(\left|\left|x^N - x^*\right|\right| \geq \rho_s\right) \leq C_1^K \exp(-C_2^K N).$$

The rest of the argument and the subsequent probabilities will be developed conditional on $\left|\left|x^N - x^*\right|\right| \leq \rho_S$. In this situation, it follows from Lemma 2 (a), that $\mathcal{I}(x^N) \subset \mathcal{I}(x^*)$. LICQ holds at (x^N) from Lemma 2 (b), and thus KKT holds and, from the definition (51) there will exist Lagrange multipliers $\lambda^N \subset \mathbb{R}^{\text{card}(I(x^*)}$ satisfying

$$\nabla_x f^N(x^N) + J^*(x^N)\lambda^N = 0.$$

It then follows, with the notation from Lemma 2 (c) that $\lambda^N = \lambda(x^N,\eta)$, where $\eta = \nabla_x f^N(x^N) - \nabla_x f(x^N)$. We also have from Lemma 2 c) that $||\eta|| \leq \delta_S$ implies that $\lambda_i^N = \lambda_i^N(x^N,\eta) > 0$, $\forall i \in I(x^*)$. In turn, from the complementarity condition at optimality for (3) we have that $\mathcal{I}(x^N) = \mathcal{I}(x^*)$. To summarize, we have that

$$\left(\left|\left|x^N - x^*\right|\right| \leq \rho_S\right) \wedge \left(\left|\left|\nabla_x f^N(x^N) - \nabla_x f(x^N)\right|\right| \leq \delta_S\right) \Rightarrow \mathcal{I}(x^N) = \mathcal{I}(x^*).$$

Using the properties of the probabilities that $\mathbb{P}(A \wedge B) \geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c)$, and that $A \Rightarrow B$ results in $P(A) < P(B)$, we obtain that

$$\begin{aligned}\mathbb{P}\left(\mathcal{I}(x^N) = \mathcal{I}(x^*)\right) \geq 1 &- \mathbb{P}\left(\left|\left|\nabla_x f^N(x^N) - \nabla_x f(x^N)\right|\right| \geq \delta_S\right) \\ &- \mathbb{P}\left(\left|\left|x^N - x^*\right|\right| \geq \rho_S\right).\end{aligned}$$

Now, using Theorem 1 and Lemma 1, we obtain that

$$\mathbb{P}\left(\mathcal{I}(x^N) = \mathcal{I}(x^*)\right) \geq 1 - \frac{C_1^g}{\rho_S^N} \exp(-C_2^g N \rho_S^2) - C_1^K \exp(-C_2^K N).$$

The conclusion follows by taking $C_1^I = 2\max\{\frac{C_1^g}{\rho_S^N}, C_1^K\}$ and $C_2^I = \min\{C_2^g \rho_S^2, C_2^K\}$. □

To simplify our subsequent analysis, we will aim to make the program (1) locally equivalent to an unconstrained optimization program based on LICQ and strict complementarity. Indeed, [A4] leads to the fact that there exists a set of columns of $J(x^*)$ whose rank is exactly $\text{card}(\mathcal{I}(x^*))$.

Notation: Given a vector $x \in X$, we denote by x^d the vector containing the components of x corresponding to indeces $J(x^*)$ and by x^e the vector containing the remaining components of x. We will use x and (x^e, x^d) interchangeably, since they are identical (modulo a permutation). As before, a superscript * denotes an optimal solution, however, for the remaining of this section the reader should be aware of the difference between x^{*e}, which refers a sub-vector of x^*, and x^{e*}, which denotes the solution of an optimization problem. The norm notation $\|\cdot\|$ refers to the Euclidian norm.

The implicit function theorem implies that there exist a twice continuously differentiable function $h : \mathcal{N}(x^{e*}) \subset \mathbb{R}^{n-\text{card}(\mathcal{I}(x^*))} \to \mathbb{R}^{\text{card}(\mathcal{I}(x^*))}$ and a neighborhood $\mathcal{N}(x^*) \subset \mathbb{R}^n$ that satisfies

$$x \in \mathcal{N}(x^*),\ g_{\mathcal{I}(x^*)}(x) = 0,\ x = \left(x^e, x^d\right) \Leftrightarrow g_{\mathcal{I}(x^*)}(x^e, h(x^e)) = 0. \tag{53}$$

This setup allows us to define the reduced stochastic program

$$\min_{x^e} f^e(x^e) := \mathbb{E}[F((x^e, h(x^e)), u)]. \tag{54}$$

It trivially follows from (53) and assumptions [A4]-[A6] that x^{*e} is a solution of this problem.

We define now a ball neighborhood of radius ρ_R of x^{*e} such that the lifting of x^e to the space X belongs to the neighborhood of x^* defined in Lemma 2; that is,

$$||x^e - x^{*e}|| \leq \rho_R \Rightarrow ||(x^e, h(x^e)) - x^*|| \leq \rho_S. \tag{55}$$

4.3 Unconstrained Stochastic Program Analysis

We also define its SAA approximation of (54),

$$\min_{x^e} f^{eN}(x^e) := \frac{1}{N} \sum_{i=1}^{N} [F^e(x^e, u_i)], \tag{56}$$

for the same samples as (3). Here, we have defined

$$F^e(x^e, u) := F((x^e, h(x^e)), u). \tag{57}$$

We note that if x^N is a solution of (3), then x^{eN} is a solution of (56). Reciprocally in a small neighborhood of x^{*e}, if x^{eN} is a solution of (56) (with the assumptions used here, it is also unique), then $x^N = (x^{eN}, h(x^{eN}))$ is a solution of (3).

We now analyze the properties of x^{eN} with increasing N, restricted to the compact set $K = \overline{\mathcal{B}(x^{e*}, \rho_R)}$. Based on the observation above, we will recover from this the properties of x^N. *Moreover, we make the entire analysis conditional on the solution x^{eN} of* (56) *belonging to* $\mathcal{B}(x^{*e}, \rho_R)$. Therefore, for the rest of this subsection we assume that $x^{eN} \in \mathcal{B}(x^{e*}, \rho_R)$.

We use the following notation:

$$L_1^N = \sup_{x^e \in K} \left|\left| \nabla_{x^e} f^e(x^e) - \nabla_{x^e} f^{eN}(x^e) \right|\right|, \tag{58}$$

$$L_3 = \sup_{\substack{x^e{}_1, x^e{}_2, x^e{}_3, x^e{}_4 \in K \\ x^e{}_1 \neq x^e{}_2,\ x^e{}_3 \neq x^e{}_4}} \frac{\left|\left| \left(\nabla^2_{x^e x^e} f^e(x^e{}_3) - \nabla^2_{x^e x^e} f^e(x^e{}_4)\right)(x^e{}_1 - x^e{}_2) \right|\right|}{||x^e{}_1 - x^e{}_2||\,||x^e{}_3 - x^e{}_4||},$$

$$\frac{1}{\sigma_0} = \sup_{x^e \in K} \left|\left| \nabla^2_{x^e x^e} f^e(x^e)^{-1} \right|\right|. \tag{59}$$

We note that L_3 and σ_0 exist following assumption [A2] and [A4]-[A6], coupled with the implicit function theorem. With this notation we can state the following results.

Lemma 4

$$\forall x^e{}_1, x^e{}_2 \in K: \quad ||\nabla_{x^e} f^e(x^e{}_2) - \nabla_{x^e} f(x^e{}_1)|| \geq \sigma_0 \,||x^e{}_2 - x^e{}_1||\,.$$

Proof First, observe that

$$\begin{aligned}
(x^e{}_2 - x^e{}_1)^T \left(\nabla_{x^e} f^e(x^e{}_1) - \nabla_{x^e} f^e(x^e{}_2)\right) &= \\
&= \int_0^1 (x^e{}_2 - x^e{}_1)^T \nabla^2_{x^e x^e} f\left(x^e{}_1 + t\,(x^e{}_2 - x^e{}_1)\right)(x^e{}_2 - x^e{}_1)\,dt \\
&\overset{(59)}{\geq} \sigma_0 \,||x^e{}_2 - x^e{}_1||^2\,.
\end{aligned}$$

By the preceding inequality and Cauchy-Schwarz inequality, we obtain that

$$\begin{aligned}
\sigma_0 \,||x^e{}_2 - x^e{}_1||^2 &\leq (x^e{}_2 - x^e{}_1)^T \left(\nabla_{x^e} f^e(x^e{}_1) - \nabla_{x^e} f^e(x^e{}_2)\right) \\
&\leq ||x^e{}_2 - x^e{}_1||\,||\nabla_{x^e} f^e(x^e{}_1) - \nabla_{x^e} f^e(x^e{}_2)||\,.
\end{aligned} \tag{60}$$

In turn, this imples that

$$||\nabla_{x^e} f^e(x^e{}_1) - \nabla_{x^e} f^e(x^e{}_2)|| \geq \sigma_0 \,||x^e{}_2 - x^e{}_1||\,,$$

which proves the claim. □

Lemma 5 *Let x^{eN} and x^{e*} be such that $\nabla_{x^e} f^{eN}(x^{eN}) = 0$ and $\nabla_{x^e} f^e(x^{e*}) = 0$. Then*

$$\left|\left| x^{eN} - x^{e*} \right|\right| \leq \frac{L_1^N}{\sigma_0}.$$

Proof We have that

$$\begin{aligned}\sigma_0 \left|\left|x^{eN} - x^{e*}\right|\right| &\leq \left|\left|\nabla_{x^e} f^e(x^{eN}) - \nabla_{x^e} f^e(x^{e*})\right|\right| \quad \text{(based on Lemma 4)} \\ &= \left|\left|\nabla_{x^e} f^e(x^{eN})\right|\right| = \left|\left|\nabla_{x^e} f^e(x^{eN}) - \nabla_{x^e} f^{eN}(x^{eN})\right|\right| \\ &\overset{(58)}{\leq} L_1^N.\end{aligned}$$

This proves the claim. □

Theorem 2 *Under the assumptions of Lemma 5 we have that*

$$f^e(x^{e*}) = f^e(x^{eN}) - \frac{1}{2}\left[\nabla_{x^e} f^e(x^{eN})\right]^T \cdot \left[\nabla^2_{x^e x^e} f^e(x^{eN})\right]^{-1} \cdot \\ \cdot \nabla_{x^e} f^e(x^{eN}) + \psi_2^N,$$

where $|\psi_2^N| \leq \Gamma \left(L_1^N\right)^3$*, for some positive constant* Γ *(independent of* n*,* x^e*).*

Proof The idea of the proof is that we do a Taylor expansion at x^{eN} for the exact mean function and use Lemma 5. We do the expansion at x^{eN} for f, and we obtain that

$$\begin{aligned}f^e(x^{e*}) =& f^e(x^{eN}) + \nabla_{x^e} f^e(x^{eN})(x^{e*} - x^{eN}) + \\ &+ \frac{1}{2}(x^{e*} - x^{eN})^T \nabla^2_{x^e x^e} f^e(x^{eN})(x^{e*} - x^{eN}) + \psi_3^N, \end{aligned} \tag{61}$$

where $\left|\psi_3^N\right| \leq L_3 \left|\left|x^{eN} - x^{e*}\right|\right|^3$.

Doing the same for the gradient, we obtain

$$0 = \nabla_{x^e} f^e(x^{e*}) = \nabla_{x^e} f^e(x^{eN}) + \nabla^2_{x^e x^e} f^e(x^{eN}) \left(x^{e*} - x^{eN}\right) + \psi_4^N, \tag{62}$$

where $\left|\left|\psi_4^N\right|\right| \leq nL_3 \left|\left|x^{e*} - x^{eN}\right|\right|^2$.

Replacing (62) in (61), we obtain

$$f^e(x^{e*}) = f^e(x^{eN}) - \frac{1}{2}\left(x^{e*} - x^{eN}\right)^T \nabla^2_{x^e x^e} f^e(x^{eN}) \left(x^{e*} - x^{eN}\right) + \psi_5^N, \tag{63}$$

with $\|\psi_5^N\| \leq L_3(1+n) \left|\left|x^{e*} - x^{eN}\right|\right|^3$.

From (62), using that $\nabla_{x^e} f^{eN}(x^{eN}) = 0$, we obtain

$$0 = -\nabla_{x^e} f^{eN}(x^{eN}) + \nabla_{x^e} f^e(x^{eN}) + \nabla^2_{x^e x^e} f^e(x^{eN}) \left(x^{e*} - x^{eN}\right) + \psi_4^N,$$

which implies

$$\left(x^{eN} - x^{e*}\right) = \left[\nabla^2_{x^e x^e} f^e(x^{eN})\right]^{-1} \left(\nabla_{x^e} f^e(x^{eN}) - \nabla_{x^e} f^{eN}(x^{eN}) + \psi_4^N\right).$$

Replacing the expression for $\left(x^{e*} - x^{eN}\right)$ in (63), collecting the residuals, and using the upper bound $1/\sigma_0$ on $\left|\left|\left(\nabla^2_{x^e x^e} f^e(x^{eN})\right)^{-1}\right|\right|$ we obtain the conclusion for $\Gamma = 2\frac{L_3(1+3n)}{\sigma_0^3}$. The proof is complete. □

We now return to the SAA interpretation, where we relate to the previous analysis by means of the notation

$$f^e(x^e) = E_u F^e(x^e, u), \quad f^{eN} = \frac{1}{N}\sum_{i=1}^{N} F(x^e, u_i).$$

It immediately follows that $\nabla_{x^e} f^e(x^{e*}) = 0$ and that $\nabla_{x^e} f^{eN}(x^{eN}) = 0$ from the optimality conditions of the stochastic problem and its SAA approximation. We define

$$\Psi(x^{eN}) = E_u F^e(x^{eN}, u) - \frac{1}{2}\left[E_u \nabla_{x^e} F^e(x^{eN}, u)\right]^T \cdot \nabla^2_{x^e x^e}\left[E_u F^e(x^{eN}, u)\right]^{-1} \cdot \cdot \left[E_u \nabla_{x^e} F^e(x^{eN}, u)\right]. \quad (64)$$

We are now in position to state our main result.

Theorem 3 *There exist $C_1(\epsilon) > 0$ and $C_2 > 0$ such that for any $a > 0$ we have that*

$$\mathbb{P}\left(\left|E_u F^e(x^{e*}, u) - \Psi(x^{eN}) - \eta^N\right| N^{-\frac{3}{2}+a} \geq \epsilon \middle| \left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R\right) \leq \leq C_1 \epsilon^{-\frac{n}{3}} N^{n\left(\frac{1}{2}-\frac{a}{3}\right)} \exp\left(-C_2 N^{\frac{a}{3}} \epsilon^{\frac{2}{3}}\right). \quad (65)$$

Here η^N is a random variable satisfying $|\eta^N| \leq \Gamma_2 (L_1^N)^3$, conditionally on $\left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R$ (and can be, for example, $\eta^N = 0$).

Note that the probability is stated conditionally on x^{eN} being sufficiently close to x^{e*}.

Proof We assume that $\left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R$, which make all of our statements conditional statements (as is the conclusion we aim to prove).

We rewrite the conclusion of Theorem 2 using the definition of Ψ to obtain that

$$E_u F^e(x^{e*}, u) = \Psi(x^{eN}) + \eta^N + \psi_2^N, \quad \left|\psi_2^N\right| \leq (\Gamma + \Gamma_2)\left(L_1^N\right)^3.$$

We thus obtain (using $0 < A < B \Rightarrow \mathbb{P}(A > \epsilon) < \mathbb{P}(B > \epsilon)$) that

$$\begin{aligned}
&\mathbb{P}\left(\left|E_u F^e(x^{e*}, u) - \Psi(x^{eN})\right| N^{-\frac{3}{2}+a} \geq \epsilon \middle| \left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R\right) \leq \\
&\mathbb{P}\left((\Gamma + \Gamma_2)\left(L_1^N\right)^3 N^{-\frac{3}{2}+a} \geq \epsilon \middle| \left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R\right) = \\
&\mathbb{P}\left(L_1^N N^{-\frac{1}{2}+\frac{a}{3}} \geq \epsilon^{\frac{1}{3}} (\Gamma + \Gamma_2)^{-\frac{1}{3}} \middle| \left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R\right) \overset{\text{Lemma 1}}{\leq} \\
&C_1^g \epsilon^{-\frac{n}{3}} (\Gamma + \Gamma_2)^{\frac{n}{3}} N^{n\left(\frac{1}{2}-\frac{a}{3}\right)} \exp\left(-\frac{C_2^g}{(\Gamma + \Gamma_2)^{\frac{2}{3}}} N^{\frac{2a}{3}} \epsilon^{\frac{2}{3}}\right).
\end{aligned}$$

Here the parameters C_1^g and C_2^g are obtained with Lemma 1 for the reduced stochastic programs (56) and (54), where the compact set K for which assumptions [A1] and [A2] hold are defined is $\left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R$. The conclusion follows after defining $C_1 = C_1^g(\Gamma + \Gamma_2)^{\frac{n}{3}}$ and $C_2 = \frac{C_2^g}{(\Gamma+\Gamma_2)^{\frac{2}{3}}}$. □

A difficulty is that, in our case, $F^e(x^e, u)$ is not twice differentiable but once differentiable, and the derivative is uniformly directionally differentiable, with bounded directional derivative.

Define the directional derivatives

$$D^i \nabla_{x^e} F^e(x^e, u) = \lim_{t \to 0, t>0} \frac{\nabla_{x^e} F^e(x^e + te_i, u) - \nabla_{x^e} F^e(x^e, u)}{t},$$

where e_i, $i = 1, 2, \ldots, n_{x^e}$, are the canonical vectors,

$$e_i = \begin{pmatrix} & & i\text{th position} & & \\ & & \downarrow & & \\ 0, 0, \ldots, 0, & & 1 & , 0 \ldots, 0 & \\ & & & & \end{pmatrix}^T.$$

Theorem 4 *The following identity holds:*

$$\nabla^2_{x^e x^e} E_u \left[F^e(x^e, u)\right] e_i = E_u \left[D^i \nabla_{x^e} F^e(x^e, u)\right].$$

Proof Using the dominated convergence theorem, we obtain that for any x^e

$$D^i E_u \left[\nabla_{x^e} F^e(x^e, u)\right] = E_u \left[D^i \nabla_{x^e} F^e(x^e, u)\right], i = 1, 2, \ldots, n_{x^e}. \tag{66}$$

Similarly, we have that $E_u \left[\nabla_{x^e} F^e(x^e, u)\right] = \nabla_{x^e} E_u \left[F^e(x^e, u)\right]$. This implies, since $E_u \left[\nabla_{x^e} F^e(x^e, u)\right]$ is differentiable, that

$$D^i E_u \left[\nabla_{x^e} F^e(x^e, u)\right] = \nabla^2_{x^e x^e} E_u \left[F^e(x^e, u)\right] e_i.$$

From the above equation and (66), the conclusion follows. □

Using Theorem 4, we obtain that

$$\begin{aligned} \Psi(x^{eN}) = E_u F^e(x^{eN}, u) - \frac{1}{2} \left[E_u \nabla_{x^e} F^e(x^{eN}, u)\right]^T \cdot \\ \cdot \left[E_u H(x^{eN}, u)\right]^{-1} \cdot \left[E_u \nabla_{x^e} F^e(x^{eN}, u)\right], \qquad (67) \\ H(x^e, u) = \left[D^1 \nabla_{x^e} F^e(x^e, u), D^2 \nabla_{x^e} F^e(x^e, u), \ldots, D^{n_{x^e}} F^e(x^e, u)\right]. \end{aligned}$$

We can now use a *different sample* $v_1, v_2, \ldots, v_n$ to estimate $\Psi(x^{eN})$ and build confidence intervals using bootstrapping. This is what we exactly propose in the next section for the more general case, namely, constrained nonlinear stochastic problems (1).

4.4 Application to the Nonlinear Stochastic Programming

In this section we state our main result that applies to nonlinear stochastic problems (1), which are assumed to satisfy the assumptions [A1]-[A6]. We write its sample average approximation (3) in the following form:

$$\begin{array}{ll} \min_x & \frac{1}{N}\sum_{i=1}^{N} F(x, u_i(\omega)), \\ \text{subject to} & g_i(x) = 0, \quad i = 1, 2, \ldots, p, \\ & g_i(x) \leq 0, \quad i = p+1, \ldots, q, \end{array} \tag{68}$$

and we define its solution by $x^N(\omega)$, and its Lagrange multipliers—if they exist–by λ^N with components λ_i^N. We define the following estimator:

$$\begin{aligned} \Upsilon(x^N, \lambda^N) = \mathbb{E}_u F(x^N, u) - \frac{1}{2} \begin{bmatrix} \mathbb{E}_u \nabla_x L(x^N, \lambda^N, u) \\ 0 \end{bmatrix}^T \cdot \\ \cdot \begin{bmatrix} \mathbb{E}_u \widetilde{H}(x^N, \lambda^N, u) & J(x^N)^T \\ J(x^N) & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \mathbb{E}_u \nabla_x L(x^N, \lambda^N u) \\ 0 \end{bmatrix}, \end{aligned} \tag{69}$$

where

$$\begin{aligned} \widetilde{H}(x^N, \lambda^N, u) &= H(x^N, u) + \sum \lambda_i \nabla_{xx}^2 g_i(x^N), \\ H(x, u) &= \left[D^1 \nabla_x F(x, u), D^2 \nabla_x F(x, u), \ldots, D^{n_x} F(x, u) \right]. \end{aligned}$$

We define by $\widehat{\Upsilon}(x^N)$ the estimator of $\Upsilon(x^N)$ using a second sample of size N (see (85) for its expression). We construct the bootstrap confidence intervals based on $\widehat{\Upsilon}(x^N)$.

The following is our main result.

Theorem 5 *Assume that [A1]-[A6] hold for the formulation* (68)*. Also let* $\widehat{\mathcal{J}}_1$ *be either the bootstrap confidence interval* $\hat{J}_1$ *(given by* (24)*) or* $\hat{J}_{1,BCa}$ *(given by* (45)*) constructed for* $\widehat{\Upsilon}(x^N, \lambda^N)$ *based on a second sample, with* λ^N *obtained from the sample average approximation of* (1)*. Then*

$$\mathbb{P}(f(x^*) \in \widehat{\mathcal{J}}_1(\alpha)) = \alpha + O(N^{-1+a})$$

for any $a > 0$*.*

Proof The proof consists of several stages.

Reduction to the framework in §4.3 We can invoke all relevant results from §4 since assumptions [A1]-[A6] hold.

Conditioning Events

Therefore, from Theorem 1 it follows that

$$\mathbb{P}(\left|\left|x^N - x^*\right|\right| \geq \rho_R) \leq C_1 \exp(-NC_2).$$

When $\left|\left|x^N - x^*\right|\right| \leq \rho_R$, it follows that $\left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R$ and the results from Section 4.3 apply. Following Lemma 3 we have that the probability that x^N has the same active set as x^* grows exponentially to 1 with N. In particular

$$\mathbb{P}\left(\mathcal{I}(x^N) = \mathcal{I}(x^*)\right) \geq 1 - C_1^I \exp(-C_2^I N).$$

We now condition on the event $\mathcal{I}(x^N) = \mathcal{I}(x^*)$ as well.

In summary, we will condition our statements on the events $\mathcal{I}(x^N) = \mathcal{I}(x^*)$ and $\left|\left|x^N - x^*\right|\right| \leq \rho_R$. From here on we assume both hold, until the final probability calculation.

Equivalence between specially constrained and unconstrained metrics

We now prove that this assumption implies that

$$\Psi(x^{eN}) = \Upsilon(x^N, \tilde{\lambda}^N), \tag{70}$$

where we have defined the special multiplier

$$\tilde{\lambda}^N := -\left(\nabla_{x_d} g_{\mathcal{I}(x^*)}(x^N)\right)^{-T} \nabla_{x_d} f(x^N). \tag{71}$$

We call it a "special" multiplier, because x^N is not a stationary point of a problem whose objective function is f; therefore, it cannot be a proper multiplier. Also note that it is a theoretical concept for analyis purposes only, since to compute it would require computing f and the integral that defines it exactly. Recall that x_d is the complement of x_e in x, as defined by (53).

Indeed, define $V^*(x) \in \mathbb{R}^n \times \mathbb{R}^{n-\mathcal{I}(x^*)}$ to be a matrix of full column rank parameterizing the nullspace of $J^*(x)$; that is, $J^*(x)V^*(x) = 0$. $V^*(x)$ as a mapping of x need not be regular, and can be obtained—conceptually, since we will never practically compute it—from the QR factorization of the square matrix $[J^{*T}(x), 0]$.

Moreover, following (53) and $\left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R$, we have that x^{eN} is a solution of (56). Also, the active set in the definition of $\Upsilon(x^N, \lambda^N)$ is the same as the one at x^*, so $J(x^N) = J^*(x^N)$.

In the following we carry out a calculation for a generic multiplier $\hat{\lambda}^N$, which can take the place of either λ^N obtained from (3) or the special multiplier defined in (71).

Following the definition of Υ, we introduce the vector $(d_x; d_\lambda)$ defined implicitly by

$$\begin{bmatrix} \mathbb{E}_u \widetilde{H}(x^N, \hat{\lambda}^N, u) & J^*(x^N)^T \\ J^*(x^N) & 0 \end{bmatrix} \begin{bmatrix} d_x \\ d_\lambda \end{bmatrix} = \begin{bmatrix} \mathbb{E}_u \nabla_x L(x^N, \hat{\lambda}^N, u) \\ 0 \end{bmatrix}, \tag{72}$$

which, in turn, results from (69) in the expression

$$\Upsilon(x^N, \hat{\lambda}^N) = \mathbb{E}_u F(x^N, u) - \frac{1}{2} \begin{bmatrix} \mathbb{E}_u \nabla_x L(x^N, \hat{\lambda}^N, u) \\ 0 \end{bmatrix}^T \begin{bmatrix} d_x \\ d_\lambda \end{bmatrix}. \tag{73}$$

Working from (72) we see that $J^*(x^N)d_x = 0$, which in turn implies the existence of a vector d_v that satisfies $d_x = V^*(x^N)d_v$. Multiplying the first row of (72) by $(V^*)^T$ and using the fact that $(V^*(x^N))^T(J^*(x^N))^T = 0$ and, subsequently, that

$$(V^*(x^N))^T \mathbb{E}_u \nabla_x L(x^N, \hat{\lambda}^N, u) = (V^*(x^N))^T \mathbb{E}_u \nabla_x F(x^N, u),$$

we obtain that

$$(V^*(x^N))^T \widetilde{H}(x^N, \hat{\lambda}^N, u) V^*(x^N) d_v = (V^*(x^N))^T \mathbb{E}_u \nabla_x F(x^N, u).$$

In turn, using these relations again in (73) together with $J(x^*)d_x = 0$, we obtain that

$$\begin{aligned}\Upsilon(x^N, \hat{\lambda}^N) = \mathbb{E}_u F(x^N, u) - \frac{1}{2} \left[(V^*(x^N))^T \mathbb{E}_u \nabla_x F(x^N, u)\right]^T \cdot \\ \cdot \left((V^*(x^N))^T \widetilde{H}(x^N, \hat{\lambda}^N, u) V^*(x^N)\right)^{-1} \cdot \\ \cdot \left[(V^*(x^N))^T \mathbb{E}_u \nabla_x F(x^N, u)\right]. \end{aligned} \tag{74}$$

We now express the quantities from (74) in terms of the quantities from (54). In terms of the notations in (53) it follows that a choice of a matrix V^* that spans the null space of J^* is

$$V^*(x^N) = \begin{bmatrix} I \\ \nabla_{x^e} h(x^e) \end{bmatrix}$$

(computed with respect to the partition (x^e, x^d)). In turn, this implies that (where we use again $f(x) \equiv \mathbb{E}F(x, u)$)

$$\begin{aligned}\left[(V^*(x^N))^T \mathbb{E}_u \nabla_x F(x^N, u)\right] &= \nabla_{x^e} f(x^N, u) + \nabla_{x^d} f(x^N, u) \nabla_{x^e} h(x^e) \\ &= \nabla_{x^e} f^e(x^e). \end{aligned} \tag{75}$$

We now look at the second-order derivatives of $f^e(x^e)$, at which point we use the definition of the multiplier $\tilde{\lambda}^N$ (71).

To that end, we note on the basis of (53) the equivalence

$$f(x^e, h(x^e)) + g(x^e, h(x^e))^T \tilde{\lambda}^N \equiv f(x^e, h(x^e)) \equiv f^e(x^e)$$

It immediately follows, on the basis of this equivalence, using the chain rule and differentiating twice, that

$$\begin{aligned}\nabla^2_{x^e x^e} f(x^e, h(x^e)) = \begin{bmatrix} I \\ \nabla_{x^e} h(x^e) \end{bmatrix}^T \nabla^2_{xx} L(x, \tilde{\lambda}) \begin{bmatrix} I \\ \nabla_{x^e} h(x^e) \end{bmatrix} + \\ + \nabla_{x^e} \left(\nabla_x L(x, \tilde{\lambda}) \overbrace{\begin{bmatrix} I \\ \nabla_{x^e} h(x^e) \end{bmatrix}}^{\downarrow} \right), \end{aligned}$$

where the symbol $\downarrow$ is used to point out to which part of the expression the differentiation is applied. The last term is then equivalent to

$$\nabla_{x^e} \left(\nabla_{x_d} L(x, \tilde{\lambda}) \overbrace{\nabla_{x^e} h(x^e)}^{\downarrow} \right),$$

which must be 0, since by the choice of $\tilde{\lambda}^N$ we have from (71) that $\nabla_{x_d} L(x, \tilde{\lambda}) = 0$. In turn, this implies that

$$\begin{aligned}\nabla^2_{x^e x^e} f(x^e, h(x^e)) &= \begin{bmatrix} I \\ \nabla_{x^e} h(x^e) \end{bmatrix}^T \nabla^2_{xx} L(x, \tilde{\lambda}) \begin{bmatrix} I \\ \nabla_{x^e} h(x^e) \end{bmatrix} \\ &= (V^*(x^N))^T \widetilde{H}(x^N, \tilde{\lambda}^N, u) V(x^N),\end{aligned}$$

the last statement following from commuting the differentiation with the $\mathbb{E}$ operator (which is legitimate because of the twice continuous differentiability of the functions involved). We have now verified that all terms in expression (74) are equal to those in the definition of Ψ, (64) which proves the claim (70).

Conditional estimate for Υ . We first bound the distance between $\tilde{\lambda}^N$ defined in (71) and the multiplier obtained by solving (3). Since the multiplier of (3) satisfies

$$\lambda^N := -\left(\nabla_{x_d} g_{\mathcal{I}(x^*)}(x^N)\right)^{-T} \nabla_{x_d} f^N(x^N),$$

we obtain that

$$\begin{aligned}\left|\left|\lambda^N - \tilde{\lambda^N}\right|\right| &= \left|\left|\left(\nabla_{x_d} g_{\mathcal{I}(x^*)}(x^N)\right)^{-T} \left(\nabla_{x_d} f^N(x^N) - \nabla_{x_d} f(x^N)\right)\right|\right| \\ &\leq \left|\left|\left(\nabla_{x_d} g_{\mathcal{I}(x^*)}(x^N)\right)\right|\right| L_1^N \leq \Gamma_L L_1^N,\end{aligned}$$

where we used (58), and, for the last inequality, the result of Lemma 2. Using the identity (74) (which holds for a generic $\hat{\lambda}^N$) and the identity (75) we obtain that there exists a Γ such that

$$|\Upsilon(x^N, \lambda^N) - \Upsilon(x^N, \tilde{\lambda}^N)| \leq \Gamma_{L2} \left|\left|\nabla_e f^e(x^{eN})\right|\right|^2 L_1^N,$$

where Γ_{L2} depends on the bound of second derivatives in ρ_R and Γ_L.

Now using Lemma 5, Assumption [A2], and the fact that $\nabla_{x^e} f^e(x^{*,e}) = 0$, which in turn implies that $\left|\left|\nabla_{x^e} f^e(x^{eN})\right|\right| = O(\left|\left|x^{*,e} - x^{eN}\right|\right|)$, we obtain that there exists a Γ_{L3} such that

$$|\Upsilon(x^N, \lambda^N) - \Upsilon(x^N, \tilde{\lambda}^N)| \leq \Gamma_{L3} \left(L_1^N\right)^3.$$

Using the last relation and (70), as well as all the events with respect to which we condition, and Theorem 3 (with $\eta^N = \Upsilon(x^N, \lambda^N) - \Upsilon(x^N, \tilde{\lambda}^N)$) we obtain that there exist $C_1(\epsilon) > 0$ and $C_2 > 0$ such that for any $a > 0$ we have that

$$\begin{aligned}\mathbb{P}\left(\left|\left|E_u F(x^*, u) - \Upsilon(x^N, \lambda^N)\right|\right| N^{-\frac{3}{2}+a} \geq \epsilon \middle| \left|\left|x^{eN} - x^{e*}\right|\right| \leq \rho_R, \mathcal{I}(x^N) = \mathcal{I}(x^*)\right) \leq \\ \leq C_1 \epsilon^{-\frac{n}{3}} N^{n\left(\frac{1}{2} - \frac{a}{3}\right)} \exp\left(-C_2 N^{\frac{a}{3}} \epsilon^{\frac{2}{3}}\right).\end{aligned}$$

We now use the following relationships among probabilities:

$$P(A) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) \leq P(A|B) + P(\bar{B}),$$

and $P(\overline{B \cap C}) \leq P(\bar{B}) + P(\bar{C})$ to obtain that

$$\begin{aligned}\mathbb{P}\left(|E_u F(x^*,u) - \Upsilon(x^N,\lambda^N)|N^{-\frac{3}{2}+a} \geq \epsilon\right) \leq \\ \leq C_1 \epsilon^{-\frac{n}{3}} N^{n\left(\frac{1}{2}-\frac{a}{3}\right)} \exp\left(-C_2 N^{\frac{a}{3}} \epsilon^{\frac{2}{3}}\right) + \\ + C_1^I \exp(-C_2^I N) + C_2^K \exp(-C_2^K N). \end{aligned} \tag{76}$$

Here we have used that the complements of the events on which we are conditioning are bounded by the results in Lemma 3 and Theorem 1.

Using the bootstrap estimate We now note that the estimator Υ (69) contains only smooth (in effect, polynomial) functions of the means of random variables. We note also that this statement does not regard variations with x, which is held fixed, but does regard functions of random variables defined at x^N, which have proper moment conditions.

Using the bootstrap theory, we obtain, with the notation of (24), that

$$\mathbb{P}(\Upsilon(x^N,\lambda^N) \leq \hat{\theta} - N^{-0.5}\hat{\sigma}\hat{y}_{1-\alpha}) = \alpha + O(N^{-1}) \tag{77}$$

$$\mathbb{P}(\Upsilon(x^N,\lambda^N) \geq \hat{\theta} - N^{-0.5}\hat{\sigma}\hat{y}_{1-\alpha}) = 1 - \alpha + O(N^{-1}) \tag{78}$$

uniformly for $\alpha \in (\eta, 1-\eta)$ and $\eta \in (0, 0.5)$. Here $\hat{\theta}$ is the empirical value of Υ (with x^N, λ^N fixed, but the mean estimated with the new sample), $\hat{\sigma}$ is the empirical standard deviation, and $\hat{y}_{1-\alpha}$ the bootstraped quantile, satisfying $\hat{K}(\hat{y}_{1-\alpha}) = 1-\alpha$. From (32), and the uniform Lipschitz property of all the terms involved in that expansion, as well as the fact that for Υ we can take $k = \infty$ in (32), we obtain that

$$|\hat{K}(\hat{y}_{1-\alpha} + \delta) - \alpha| \leq C\delta + O(N^{-\frac{k+1}{2}}), \tag{79}$$

for some fixed C and k.

To simplify notation, we denote $\mathbf{E}_u F(x^*,u) = \Psi^*$, and $\hat{\Upsilon} \equiv \hat{\Upsilon}(x^N,\lambda^N)$. Also let $\hat{w}_\alpha$ generically denote either $\hat{y}_\alpha$ or $\hat{z}_{BCa,\alpha}$. Keep in mind that $|\hat{w}_\alpha - \hat{y}_\alpha|$ is either 0 or $O(N^{-1})$ (from (44)). In (76) fix $\epsilon = 1$.

Using that $P(A) \leq P(A|B) + P(\bar{B})$, we obtain that

$$\begin{aligned}&\mathbb{P}(\Psi^* \geq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{w}_{1-\alpha}) \leq \\ &\mathbb{P}\left(\Psi^* \geq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{w}_{1-\alpha} \,\middle|\, \left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \leq N^{-\frac{3}{2}+a}\right) \\ &\qquad + \mathbb{P}\left(\left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \geq N^{-\frac{3}{2}+a}\right) \leq \\ &\mathbb{P}\left(\hat{\Upsilon} \geq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{w}_{1-\alpha} - N^{-\frac{3}{2}+a}\right) + \mathbb{P}\left(\left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \geq N^{-\frac{3}{2}+a}\right) \overset{(44)}{\leq} \\ &\mathbb{P}\left(\hat{\Upsilon} \geq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{y}_{1-\alpha} + O(N^{-\frac{3}{2}}) - N^{-\frac{3}{2}+a})\right) \\ &\qquad + \mathbb{P}\left(\left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \geq N^{-\frac{3}{2}+a}\right) \overset{(79),(78),(76)}{\leq} \\ &1 - \alpha + O(N^{-1}) + O(N^{-\frac{k+1}{2}}) + CN^{-1+a}\frac{1}{\hat{\sigma}} + O(N^{n\left(\frac{1}{2}-\frac{a}{3}\right)} \exp(-C_2' N^{\frac{a}{3}})),\end{aligned}$$

where in the exponential terms we incorporated all the $\exp(-N)$ term in (76) in the last term. After analyzing all leading terms (for fixed, but arbitrary $0.5 > a > 0$), we obtain that

$$\mathbb{P}(\Psi^* \geq \hat{\theta} - N^{-0.5}\hat{\sigma}\hat{w}_{1-\alpha}) \leq 1 - \alpha + O(N^{-1+a}). \tag{80}$$

A similar analysis, working now with (77), leads to the opposite bound as well. Using again that $P(A) \leq P(A|B) + P(\bar{B})$, we obtain that

$$\begin{aligned}
&\mathbb{P}(\Psi^* \leq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{w}_{1-\alpha}) \leq \\
&\mathbb{P}\left(\Psi^* \leq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{w}_{1-\alpha} \,\middle|\, \left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \leq N^{-\frac{3}{2}+a}\right) \\
&\qquad\qquad + \mathbb{P}\left(\left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \leq N^{-\frac{3}{2}+a}\right) \leq \\
&\mathbb{P}\left(\hat{\Upsilon} \leq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{w}_{1-\alpha} + N^{-\frac{3}{2}+a}\right) + \mathbb{P}\left(\left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \geq N^{-\frac{3}{2}+a}\right) \overset{(44)}{\leq} \\
&\mathbb{P}\left(\hat{\Upsilon} \leq \hat{\theta} - N^{-\frac{1}{2}}\hat{\sigma}\hat{y}_{1-\alpha} + O(N^{-\frac{3}{2}}) + N^{-\frac{3}{2}+a})\right) \\
&\qquad\qquad + \mathbb{P}\left(\left|\left|\Psi^* - \hat{\Upsilon}\right|\right| \geq N^{-\frac{3}{2}+a}\right) \overset{(79),(77),(76)}{\leq} \\
&\alpha + O(N^{-1}) + O(N^{-\frac{k+1}{2}}) + CN^{-1+a}\frac{1}{\hat{\sigma}} + O(N^{n\left(\frac{1}{2}-\frac{a}{3}\right)}\exp(-C_2' N^{\frac{a}{3}})),
\end{aligned}$$

In turn, this results in

$$\mathbb{P}(\Psi^* \leq \hat{\theta} - N^{-0.5}\hat{\sigma}\hat{w}_{1-\alpha}) \leq \alpha + O(N^{-1+a}).$$

Using (80), we obtain that

$$\begin{aligned}
\mathbb{P}(\Psi^* \leq \hat{\theta} - N^{-0.5}\hat{\sigma}\hat{w}_{1-\alpha}) &= 1 - \mathbb{P}(\Psi^* \geq \hat{\theta} - N^{-0.5}\hat{\sigma}\hat{w}_{1-\alpha}) \\
&\geq 1 - (1-\alpha) + O(N^{-1+a}) = \alpha + O(N^{-1+a}).
\end{aligned}$$

From the definitions (24) and (45), the last two equations prove the claim.□

5 Numerical simulations

5.1 Test problems

The numerical performance of the estimator $\hat{\Upsilon}$ is studied with two test problems. Since the validation of the nominal coverage requires extensive simulations, we chose small problems to keep the computational cost tractable. The first problem, which we call *TOY*, is a one-dimensional unconstrained problem given by

$$\min f(x) := E_u(x-u)^4, \tag{81}$$

where $u \sim U(0,1)$, with the optimal solution $x^* = 0.5$ and optimal value $f(x^*) = 0.0125$. The SAA problem is

$$\min f^N(x) := \frac{1}{N}\sum_{i=1}^{N}(x-u_i)^4, \tag{82}$$

where u_i, $i = 1,\ldots,N$, are random i.i.d realizations drawn from $U(0,1)$. Due to the smoothness of the function $F(x,u)$ in the casting of (1), assumptions [A0]-[A6] hold in this case.

The second problem is a two-stage stochastic optimization problem with recourse, which we call *PROB2Q*. Even though stochastic programming problems with recourse do not necessarily fit the framework of Section 4.4, we present the behavior of bootstrap confidence intervals and show numerical evidence that they are superior to "normal" confidence intervals. PROB2Q is a modification of a oil refinery model [10] that describes the weekly production process of a refinery that buys crude oil from two sources and has to supply two clients with gasoline and heating fuel. We added quadratic costs both in the first stage (costs of buying crude oil) and in the second stage ("penalty" costs induced by the incapacity of satisfying demand due to unforseen events u_1 and u_2 in determining the demand). PROB2Q problem takes the following form

$$\min f(x_1,x_2) := \frac{5}{2}x_1^2 + \frac{5}{2}x_2^2 + 7.4x_1 + 2.4x_2 + E_{u_1,u_2}Q(x_1,x_2;u_1,u_2), \tag{83}$$

where

$$\begin{array}{rll} Q(x_1,x_2;u_1,u_2) = \min_{y_1,y_2} & \frac{1}{2}(y_1^2+y_2^2) - 2y_1 - 2y_2 & \\ \text{s.t.} & y_1 \geq 20 + u_1 - (2x_1+6x_2) & \\ & y_2 \geq 10 + u_2 - (3x_1+3x_2). & \end{array} \tag{84}$$

Here $u_1 \sim U(-10,10)$ and $u_2 \sim U(-5,5)$. *PROB2Q* has a unique optimal solution at $x^* = (x_1,x_2) = (3,0)$ and an optimal value $f(x^*) = 35.1$. As we discussed following the statement of assumption [A2], at the moment we cannot state that our theory applies for the two-stage stochastic programming case, but it is an important case to study empirically, in particular, because [A2] cannot be ensured to hold (all the other assumptions can be proved to hold here). Moreover, following the theory from [2], it can be shown that $F(x,u)$ is twice directionally differentiable when mapping (84) in the framework (1). *Therefore the estimator* (69) *does exist*, though we cannot state that our main result Theorem 5 holds.

5.2 Bootstrap Numerical Method

In this section we present the bootstrap methodology we used to compute confidence intervals based on the estimator Υ given by (69). Recall that a sample $(u_1,u_2,\ldots,u_N)$ was used to pinpoint a point x_N at which Υ is defined.

Since Υ is a function of moments, a second sample $(v_1, v_2, \ldots, v_N)$ is needed to obtain the estimator

$$\widehat{\Upsilon} = \widehat{\Upsilon}(v_1, v_2, \ldots, v_N) = \frac{1}{N}\sum_{i=1}^{N} F(x_N, v_i) - \frac{1}{2}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \nabla_x L(x^N, \lambda^N, v_i) \\ 0 \end{bmatrix}^T \cdot \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \widetilde{H}(x^N, \lambda^N, v_i) & J(x^N)^T \\ J(x^N) & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} \nabla_x L(x^N, \lambda^N, v_i) \\ 0 \end{bmatrix}, \quad (85)$$

which is bootstrapped to obtain the confidence intervals.

In our preliminary numerical tests we observed that the BCa method is superior to Hall's method for small-sized samples. In our opinion this behavior is caused by large errors present in the variance estimate $\hat{\sigma}^2$ used by Hall's method. We now present the methodology of constructing bootstrap α-level confidence intervals using the BCa method. A short introduction to BCa method was given in Section 3.1.2; a detailed discussion can be found in [6].

Procedure for computing the BCa α-level one-sided confidence intervals $\hat{J}_{1,BCa}(\alpha)$ for Υ

Bootstrapping $\mathcal{V} = (v_1, v_2, \ldots, v_N)$

1. Draw with replacement B samples from $\mathcal{V}$, namely $\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_B$.
2. For each $b = 1, \ldots, B$ evaluate $\widehat{\Upsilon}_b^B = \widehat{\Upsilon}(\mathcal{V}_b)$.
3. Compute $\widehat{\Upsilon}^B = \frac{1}{N}\sum_{b=1}^{B} \widehat{\Upsilon}_b^B$.

Computing the acceleration $\hat{a}$ using jackknife

4. For each $i = 1, \ldots, N$ evaluate $\widehat{\Upsilon}_i^J = \widehat{\Upsilon}(\mathcal{V}_i^J)$, where $\mathcal{V}_i^J = (v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_N)$.
5. Compute $\widehat{\Upsilon}^J = \frac{1}{N}\sum_{i=1}^{N} \widehat{\Upsilon}_i^J$.
6. Compute $\hat{a} = \sum_{i=1}^{N} (\widehat{\Upsilon}^J - \widehat{\Upsilon}_i^J)^3 / \left(6\left(\sum_{i=1}^{N} (\widehat{\Upsilon}^J - \widehat{\Upsilon}_i^J)^2\right)^{3/2}\right)$.

Bias correction and skewness adjustment

7. Compute $\hat{m} = \Phi^{-1}(\frac{1}{B}\text{card}\{\widehat{\Upsilon}_b^B < \widehat{\Upsilon} : b = 1, \ldots, B\})$ and $z_\alpha = \Phi^{-1}(\alpha)$.
8. Compute $\alpha_c = \Phi\left(\hat{m} + \frac{z_\alpha + \hat{m}}{1 - \hat{a}(z_\alpha + \hat{m})}\right)$.
9. Obtain the BCa quantile $\hat{v}_{BCa,\alpha} = \min\left\{y : \frac{1}{B}\text{card}\{\widehat{\Upsilon} \le y : b = 1, \ldots, B\} \ge \alpha\right\}$.

Return $\hat{J}_{1,BCa} = (-\infty, \hat{v}_{BCa,\alpha})$.

Fig. 1 Steps taken in the computation of the one-sided α-level confidence interval using the BCa method. The procedure can be applied to any Υ that can be expressed as a function of moments.

The computation of $\widehat{\Upsilon}$ requires the evaluation of the second-stage recourse objective, its gradient, and the directional derivative of the gradient for each sample v_i. While it is simple to get the value of the objective and the gradient by solving the second-stage problem, the directional derivative is a more computationally intensive operation. In the numerical experiments presented here we obtained the directional derivative by evaluating analytical expressions that we have been able to derive because of the small dimensionality of the test problems. In general, another optimization problem has to be solved to get the directional derivative of the gradient [2, Theorem 5.53].

It is crucial to the numerical performance of the method presented in Figure 1 to observe that the evaluation of the second-stage recourse objective, its gradient, and the directional derivative of the gradient can be done only once for each sample (a total $2N$ optimization problems) when initially computing $\widehat{\Upsilon}$ and reused both in the bootstrap and jackknife phases when computing $\widehat{\Upsilon}_b^B$'s and $\widehat{\Upsilon}_i^J$'s. Therefore, the computational cost of the method increases only linearly with B.

5.3 Performance of bootstrapping

In this section we present the numerical order of bootstrap for the estimator $\widehat{\Upsilon}$ for the TOY and PROB2Q problems. We also compare the performance of $\widehat{\Upsilon}$ with that of the "uncorrected" SAA estimator $f(x^N) = E_u F(x^N, u)$, where, as before, x^N is the solution of the SAA problem based on the first sample $u_1, \ldots, u_n$. The second sample $v_1, \ldots, v_N$ is used for bootstrapping $\widehat{f(x_N)} = \frac{1}{N}\sum_{i=1}^{N} F(x^N, v_i)$. We call this simply the SAA estimator.

We applied the BCa bootstrap methodology presented in Section 5.2 to both $\widehat{\Upsilon}$ and $\widehat{f(x_N)}$. The coverage $\mathbb{P}(f(x^*) \in \hat{J}_{1,BCa}(\alpha))$ of the CIs of the abovementioned estimators was computed by simulation, as follows. We first computed a large number (90 thousand for the TOY problem and 200 thousand for the PROB2Q problem) confidence intervals $\hat{J}_{1,BCa}(0.05)$ and $\hat{J}_{1,BCa}(0.95)$ for various sample sizes N. We then approximated the coverage by the percentage of confidence intervals that contained the true optimal value.

Table 1 Coverages of one-sided 5% (left columns) and 95% (right columns) CIs for the TOY problem. For the statistic Υ we also show coverages for the normal studentized CIs.

	Υ				$f(x^N)$	
n	Normal CIs		BCA CIs		BCA CIs	
3	0.0634	0.5071	0.0291	0.4764	0.0201	0.2613
5	0.0373	0.5965	0.0385	0.6314	0.0217	0.4990
10	0.0243	0.7264	0.0428	0.8011	0.0280	0.7515
15	0.0224	0.7889	0.0458	0.8653	0.0311	0.8387
20	0.0224	0.8268	0.0466	0.9013	0.0324	0.8764
25	0.0235	0.8471	0.0466	0.9173	0.0341	0.8971
30	0.0241	0.8624	0.0467	0.9286	0.0342	0.9099
40	0.0245	0.8798	0.0471	0.9401	0.0365	0.9243
50	0.0273	0.8934	0.0478	0.9445	0.0372	0.9293
100	0.0312	0.9152	0.0488	0.9497	0.0410	0.9377
200	0.0357	0.9262	0.0495	0.9499	0.0414	0.9418
400	0.0394	0.9358	0.0491	0.9492	0.0434	0.9432

For a given sample size N, we define the coverage error as $e_N = |\mathbb{P}(f(x^*) \in \hat{J}_{1,BCa}(\alpha)) - \alpha|$, $\alpha \in \{0.05, 0.95\}$. The e_N error should be reduced to zero with N at an $O(N^{-1+a})$ rate, $a > 0$, according to Theorem 5.

We use a robust linear regression for $e_N = \beta N^{-\gamma}$ to compute the "numerical" order of correctness of bootstrapping applied to the estimator $\hat{\Upsilon}$ defined

by Theorem 5 and SAA estimator $f(x^N)$. For $\hat{\Upsilon}$ we also computed the normal studentized CIs (9) and compared their accuracy with that of the bootstrap CIs.

Table 2 Coverages of one-sided 5% (left columns) and 95% (right columns) CIs for the PROB2Q problem. For the statistic Υ we also show coverages obtained by using normal studentized CIs.

	Υ				$f(x^N)$	
n	Normal CIs		BCA CIs		BCA CIs	
3	0.0140	0.6572	0.0705	0.5370	0.1208	0.6781
5	0.0095	0.7522	0.0887	0.7265	0.0917	0.7983
10	0.0084	0.8336	0.0720	0.8730	0.0783	0.9029
20	0.0091	0.8749	0.0579	0.9317	0.0670	0.9245
30	0.0112	0.8885	0.0537	0.9416	0.0625	0.9341
40	0.0132	0.8979	0.0525	0.9451	0.0605	0.9554
50	0.0149	0.9039	0.0515	0.9465	0.0589	0.9560
100	0.0220	0.9177	0.0508	0.9481	0.0562	0.9543
200	0.0292	0.9263	0.0505	0.9484	0.0544	0.9525
400	0.0344	0.9343	0.0492	0.9500	0.0520	0.9530

The order of correctness for the confidence intervals obtained by bootstrap for the estimator Υ, defined in this paper, was found to be better than the term $O(N^{-1+a})$ predicted by Theorem 5: 1.67(0.69) and 1.59(1.13) in the case of the PROB2Q problem (for $\alpha = 0.05$ and $\alpha = 0.95$, respectively) and 2.11(1.14) in the case of the TOY problem for $\alpha = 0.95$. Here we show in parentheses the result for the classical confidence intervals of the SAA estimator $f(x^N)$ that do not use bootstrap. For $\alpha = 0.05$ the order of correctness is 0.82, and its departure from 1 most likely is caused by the large negative skewness of the distribution of Υ Note that the order is larger by precisely 0.5 compared with the classical estimator, in any case (0.32, see Figure 2).

The correctness of confidence intervals based on the statistic Υ we proposed in this paper is superior to the correctness of the confidence intervals based on the classical SAA estimate $f(x^N)$, as can be seen in Figure 2 and Figure 3. The correctness of the confidence intervals for Υ constructed by bootstrapping is always at least one order ($O(N^{-0.5})$) better than that of $f(x^N)$.

When comparing the quality of the normal studentized CIs with that of the BCa CIs (both applied to the statistic Υ) shown in Table 1 and Table 2, one can easily see that the performance of bootstrap is better than that of studentized CIs, in terms of both accuracy at small samples and asymptotic order of correctness.

6 Conclusions

We have presented a new, bootstrap-based approach for creating statistical estimators of the optimal value of stochastic programs for which the coverage

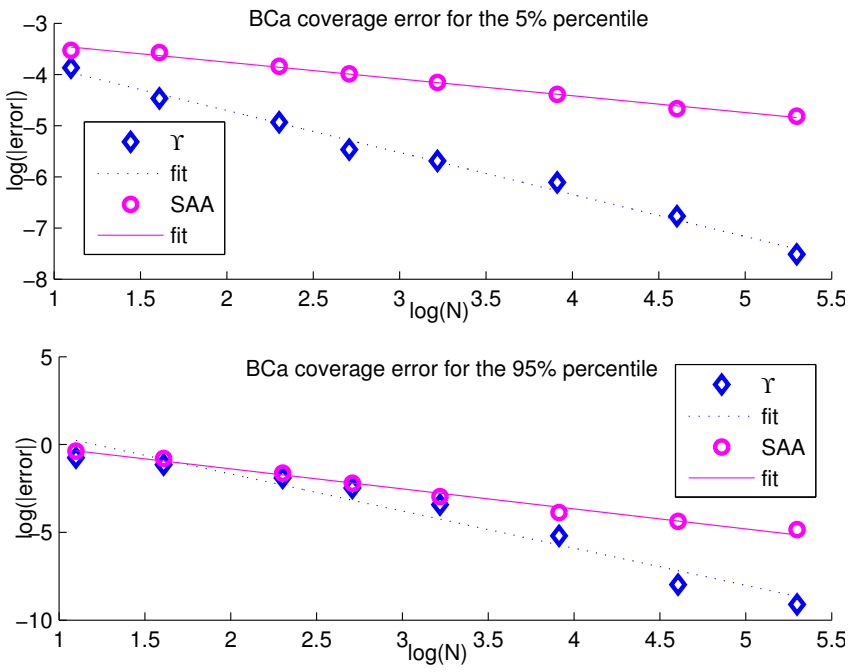

Fig. 2 Errors e_N ($N = \{3, 5, 10, 15, 25, 50, 100, 200\}$) of the BCa coverages from Table 1 (TOY problem) plotted against a robust linear regression fit for $\beta N^{-\gamma} = e_N$. The bootstrap order given by γ is the slope of the linear regression and is found to be 0.82 for Υ and 0.32 for $f(x^N)$ in the case of the 5% percentile interval $\hat{J}_{1,BCa}(0.05)$ (top subplot). The bootstrap order is 2.11 for Υ and 1.14 for $f(x^N)$ in the case of the 95% confidence interval $\hat{J}_{1,BCa}(0.95)$ (bottom subplot).

probability of the confidence intervals converge faster than the standard estimates to their nominal values. In turn, this allows a more reliable uncertainty estimate for lower sample sizes. The latter feature is essential in applications where sampling is extremely expensive, such as in stochastic unit commitment for power grid management, where the samples from the state of the atmosphere need to be produced [4].

Under some conditions about the regularity of the stochastic program, we prove that our estimates have an almost $O(N^{-1})$ coverage probability compared with standard estimates that have error $O(N^{-0.5})$. We point out, in addition, that our analytical framework allows for the evaluation of the asymptotics of the coverage probabilities, which is valuable even for standard estimates (and new, to our knowledge). The good convergence properties are demonstrated with two numerical examples.

We make several assumptions that limit the generality of our approach. We regard these assumptions as important tradeoffs as we attempt to create analytical tools to analyze the asymptotics of the coverage of confidence inter-

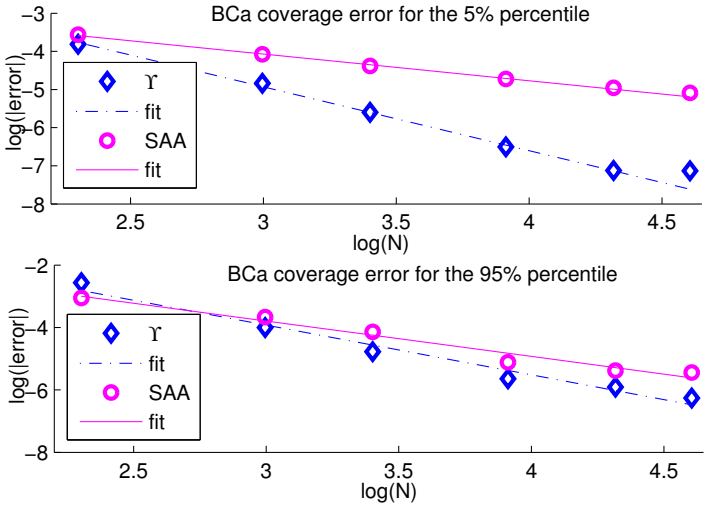

Fig. 3 Errors e_N ($N = \{10, 20, 30, 40, 50, 75, 100\}$) of the BCa coverages from Table 2 (PROB2Q problem) plotted against a robust linear regression fit for $\beta N^{-\gamma} = e_N$. The last two rows from Table 2 (corresponding to $N = 200$ and $N = 400$) were not included in the regression and in the plots since the accuracy of the coverage is likely to be affected by the simulation noise for these large values of N. The bootstrap order given by γ is the slope of the linear regression and is found to be 1.67 for Υ and 0.69 for $f(x^N)$ in the case of the 5% percentile interval $\hat{J}_{1,BCa}(0.05)$ (top subplot). The bootstrap order is 1.59 for Υ and 1.13 for $f(x^N)$ in the case of the 95% confidence interval $\hat{J}_{1,BCa}(0.95)$ (bottom subplot).

vals. Perhaps the most limiting at this moment is that the first-stage nonlinear program has a Lipschitz, twice-differentiable objective with Lipschitz second derivatives after accounting for the second-stage variable. In particular, this makes our analysis not immediately applicable to two-stage stochastic programming with inequality constraints, even as empirically the results seem to hold for the two-stage stochastic program we tested. Our analysis would hold for some approximations such as some smoothing of constraint effects. Nevertheless, relaxing this and some of the other assumptions is an important future endeavor.

Moreover, while superior asymptotically, our solution needs *two* samples. The difficulty comes from the fact that we have found no way to analyze statistical estimates based on resampling the first sample. Our approach has been to construct an estimator that involves only smooth functions of the mean of a distribution, and then to invoke results from [8]. Unfortunately, we have not yet found a way to cast the optimal value in this fashion directly in terms of finite-dimensional distributions for which the results in [8] apply. We could repeat the reduction approach in §4.3 without difficulty, and then state that x^N converges exponentially to a neighborhood of x^* where $x^{eN} = \arg\min \mathbb{E}_u[F^N((x^e, h(x^e), u)]$. Then x^{eN} is a function of the mean defined over some infinite dimensional space, which exhibits some smoothness

in a functional derivative sense. Unfortunately, the results in [8] apply only for random variables in finite-dimensional space. We add, however, that our endeavor in bootstrapping the first sample has indicated empirically substantially more variability and lower asymptotic quality than in the two-sample approach, but it is hard to say at this point how generic this is, partly because of the very high validation cost. Whether bootstrapping with one sample can work for stochastic programming—in the sense of superior convergence estimates—and how to attack this problem analytically for one sample is an intriguing question for future research.

Acknowledgments

Mihai Anitescu is grateful to Steve Lalley and Michael Stein for discussions on large deviations and convergence concepts. This work was supported by U.S. Department of Energy, under Contract DE-AC02-06CH11357.

References

1. Billingsley, P.: Probability and measure, 3rd ed. Wiley-Interscience, New York, (1995)
2. Bonnans, J., Shapiro, A.: Perturbation analysis of optimization problems. New York. Springer Verlag (2000)
3. Casella, G., Berger, R.L.: Statistical Inference. Duxbury Press (1990)
4. Constantinescu, E.M., Zavala, V.M., Rocklin, M., Lee, S., Anitescu, M.: A computational framework for uncertainty quantification and stochastic optimization in unit commitment with wind power generation. IEEE Transactions on Power Systems **26**(1), 431–441 (2010)
5. Efron, B.: The Jackknife, the Bootstrap and Other Resampling Plans. SIAM (1982)
6. Efron, B., Tibshirani, R.J.: An Introduction to the Bootstrap. Chapman & Hall (1993)
7. Hall, P.: Inverting an Edgeworth expansion. The Annals of Statistics **11**, 569–576 (1983)
8. Hall, P.: Theoretical comparison of bootstrap confidence intervals. The Annals of Statistics **16**, 927–953 (1988)
9. Hall, P.: The Bootstrap and Edgeworth Expansion. Springer-Verlag, New York (1992)
10. Kall, P., Wallace, S.W.: Stochastic Programming, 2nd edn. John Wiley & Sons, Chichester, UK (1994)
11. Nocedal, J., Wright, S.: Numerical Optimization. Springer Verlag (1999)
12. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory. MPS/SIAM Series on Optimization 9, Philadelphia, PA (2009)